

B E A N

Construction of
Educational and Personnel
TESTS

5387

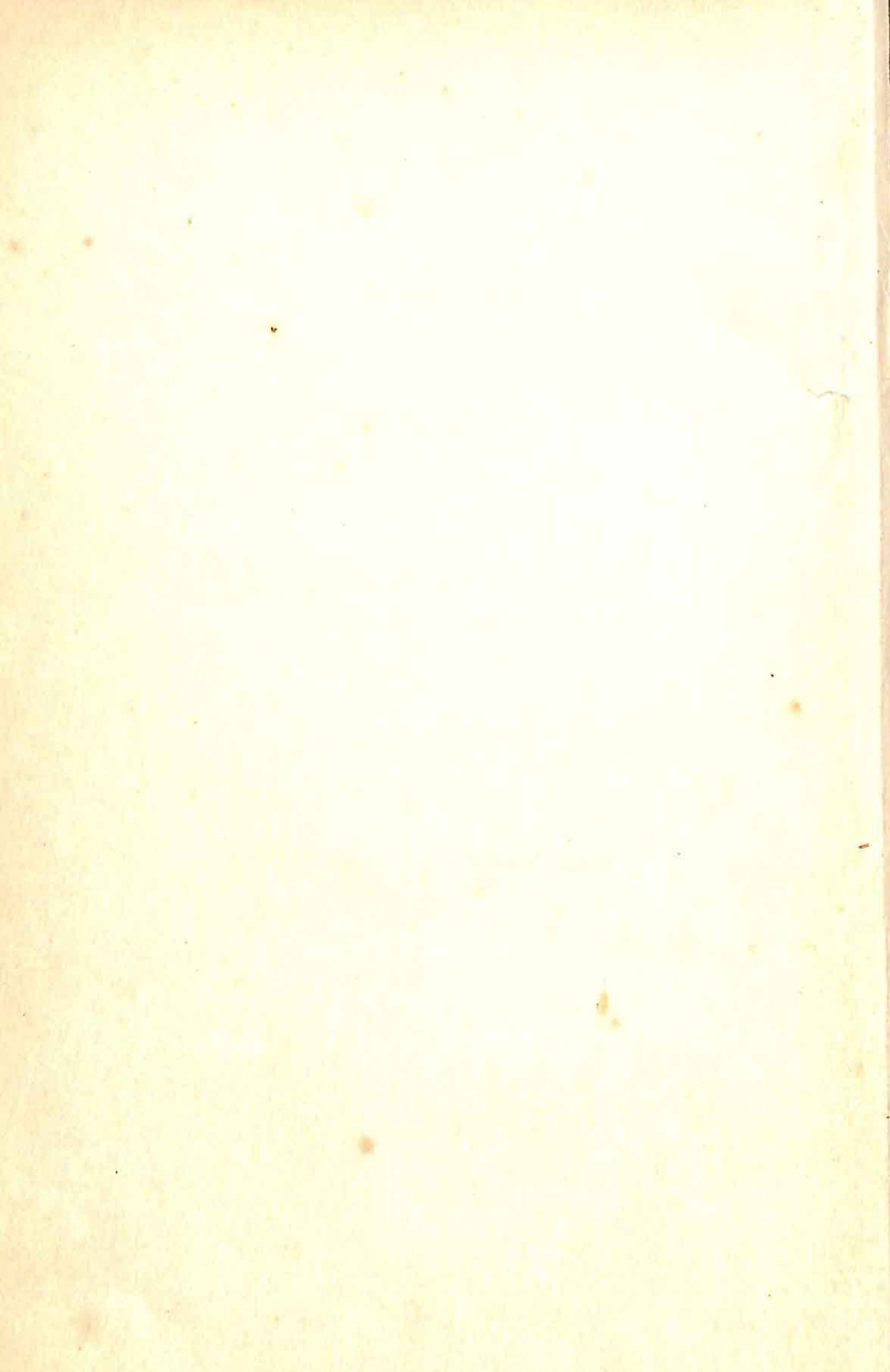


Ed. test

34126
B 367



Construction of
Educational and Personnel
TESTS



Construction of
Educational and Personnel
TESTS

KENNETH L. BEAN, PH.D.

Professor of Psychology
Baylor University

McGraw-Hill Book Company, Inc.

NEW YORK TORONTO LONDON

1953

371.26
BEA

CONSTRUCTION OF EDUCATIONAL AND PERSONNEL TESTS

Copyright, 1953, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

Library of Congress Catalog Card Number: 53-5159

591

Bureau Ednl. Psy. Research
DAVID HARE TRAINING COLLEGE
Dated.....15.3.55
Accs. No. 591

PREFACE

THROUGH six years of experience as a consulting psychologist for a state public personnel agency and nearly six years of teaching and clinical experience in psychology, the author has gathered data about and techniques of working with source materials to convert them into suitable tests for various purposes. This experience has included cooperation with experts in numerous occupational areas, including skilled and semiskilled trades, office procedures, professional social work, and law enforcement. The experience has also included investigations for the purpose of designing, writing, and compiling appropriate employment tests. Out of the rather disorganized materials and techniques the author has sifted the worth-while ideas, practices, and principles which appear most useful to those who must almost daily employ them.

Outside the academic fields there are two groups, rapidly increasing in size, which need a working knowledge of test construction. One of these includes examiners in Federal, state, county, and municipal civil service agencies. Many of such examiners have had courses in tests and measurements, but some have been promoted from clerical positions without special training in tests and measurements. Very few have had adequate training in the construction of tests. Inspection of a large sample of examinations for various occupations secured through an exchange agency has convinced the author that there is an evident need for an in-service training manual for the majority of civil service examiners.

The second of these nonacademic groups includes research workers engaged in test development, both in industry and in other agencies concerned with personnel selection and vocational guidance. Most of such investigators have adequate working knowledge of existing tests and statistical procedures. But industry requirements are continually changing; new tests must be continually devised and old ones both revalued and changed. Competence in test construction is simply a new basis for personnel directors to evaluate both their own proficiency and the proficiency of those who work under their direction.

Two overlapping categories, measures of aptitude and measures

of achievement, are frequently confused by the administrators of personnel and by educational supervisors. The author has, perhaps, been guilty of reversing the usual emphasis in his eagerness to present the techniques of how to make good tests rather than to discuss the use of tests already prepared by others.

The present volume can serve as a text in a college course, provided it is supplemented by outside readings in statistics, development of intelligence tests, and other areas somewhat outside its scope, especially if the materials in Appendix B are made available. It will be most useful to the person who is actually engaged in constructing tests. It can be read by teachers having limited knowledge of statistics and can still be applied with benefit in their courses. The material should be of special help to many college instructors, especially those of limited experience. In-service training for examiners in public and private personnel could include development of skills through practice of the principles developed in these chapters.

Never should the reader consider this volume alone as constituting adequate preparation for research in test development in any field. Thorough acquaintance with the existing literature on testing and with statistical methods are absolute essentials to a real contribution to test research. This volume is designed primarily as a handbook of basic principles and their application to specific problems. If it stimulates efforts to improve essential, practical skills, it will have served its primary purpose.

The writer wishes to express his appreciation of the valuable constructive criticism offered by Dr. F. F. Burtchett, University of Houston, the interesting data furnished by H. L. Trites, Veterans' Administration, Waco, Tex., and the assistance in research given by Mrs. Loyd Taylor, Baylor University, Waco, Tex.

Kenneth L. Bean

CONTENTS

<i>Preface</i>	v
CHAPTER 1. <i>Introduction</i>	1
Uses and Abuses of Tests	1
Practical Applications	2
Criticisms of Examination Techniques	3
Goals and Objectives	5
Classification of Tests	8
Definition of Concepts Considered	11
CHAPTER 2. <i>Planning the Examination as a Whole</i>	16
Defining the Purpose of the Test	17
Problems of Test Administration	18
The Scoring Problem	22
Weighting Scores on Composite Tests	24
Analysis of Content of Course for Tests of Achievement	26
Job Analysis as a Preliminary to Job Testing	27
Tentative Outline of the Proposed Test	31
CHAPTER 3. <i>Converting Material into Objective-test Items</i>	37
Available Sources	38
Alternative-response Items	43
Multiple-choice Items	52
Completion Test Items	75
Matching Items	80
Rare Test Forms	83
CHAPTER 4. <i>Special Problems in Objective-test Construction</i>	88
An English-usage Test for Clerical Workers	88
Employment Tests for Illiterates and Near Illiterates	91
Instructions for Tests	94
Reading Comprehension	96
Vocabulary	103
CHAPTER 5. <i>To "Essay" or Not to "Essay"</i>	107
Critique of Tests	107
Advantages of Essay Questions	108
Disadvantages of Essay Questions	111

Test Forms and Examples	116
Approaching Reasonable Objectivity in Scoring	121
Making a Key for Scoring	124
CHAPTER 6. <i>Performance Tests</i>	128
Why Some Tests Are Not Written Tests	128
Standardization of Material and Procedure	131
Methods of Rating	138
CHAPTER 7. <i>Review and Tryout of a Test</i>	142
Subject-matter Consultants	143
Technical Review from the Construction Viewpoint	150
Item Analysis	152
CHAPTER 8. <i>Validity, Reliability, and Standardization</i>	159
The Criterion	160
Reliability	165
Norms	169
Research	176
APPENDIX A. <i>Sample Problem in Test Construction and Solution</i>	179
APPENDIX B. <i>Example of Comparative Answers for a Performance-test Problem</i>	188
APPENDIX C. <i>Example of Performance and Scoring of a Nondirective Performance Test</i>	206
APPENDIX D. <i>Collateral Readings</i>	224
Index	229

INTRODUCTION

THOROUGH knowledge of the standardization, content, administration, scoring, and interpretation of a wide variety of standardized tests does not necessarily carry with it skill in test construction. To teachers at any level in our educational program, such skill is absolutely essential if the attainment of academic objectives is to be accurately and fairly evaluated. Yet comparatively few on the faculties of colleges and universities have had any specialized training in preparing tests, and the training of elementary school and high school teachers is seriously deficient in this respect. An undergraduate major in psychology or education does not often include instruction in this field. Only a few leading universities are now listing courses primarily or partially for instruction in test construction.

Even the better psychological treatises on tests and measurements pass lightly over the problem of constructing good tests and devote their space almost entirely to problems of administration, scoring, interpretation, norms, reliability, validity, and application of existing tests. Necessary as statistical data may be, this aspect of test construction is not the whole story. Only a very meager literature can be found which is helpful in a practical way in the conversion of material into good test items. Probably none of the treatises have brought together the widely scattered research studies in such a way that the results can be adequately applied in a variety of practical situations. Few present-day writers of recognized tests have incorporated into their test manuals any detailed account as to how they obtained or refined their ideas.

Uses and Abuses of Tests. College students frequently complain about the kinds of examinations their professors give. "I never know how to study for his tests" is a much overworked expression on many a university campus. Some of these complaints are no doubt mere

rationalizations for failure. It is easy for a student to project his own inadequacies upon his instructor and to give himself credit for any successes he may achieve. Perhaps that is what two sophomores were doing as they stood in front of the bulletin board on which grades were posted. "I made A," remarked one. "He gave me a D," commented the other. Yet the criticisms of examinations in many courses made by the more serious, more capable students are often too valid to be ignored.

Civil service examinations, Federal, state, county, and municipal, often receive and deserve unfavorable comment from the best-qualified applicants on the grounds that the examinations are too academic, are not practical, do not relate to the job, and that some questions are ambiguous. There are numerous outstanding exceptions that are for the most part, if not entirely, above reproach. Yet there exist agencies over the country which have borrowed from each other, sometimes through a test-exchange service, some of the most hackneyed questions ever written. Many tests currently employed cover unimportant points and disregard generally accepted principles of item or test construction. Here, again, many complaints are mere rationalizations, but one cannot deny that there is a need for new ideas in examinations used by public personnel agencies.

Practical Applications. To provide a working knowledge of basic principles in constructing and improving tests that can be used from day to day in a variety of situations is a complicated task. To do it successfully, the reader must be stimulated to a constructively critical attitude in the inspection of any attempt to measure human behavior. Specifically, the practical user must be inspired to practices leading to the development of skills in designing, compiling, writing, and trying out tests in order to achieve the most reliable possible measure of the abilities or achievements that they are intended to evaluate. Surprisingly few treatises in technical psychology dealing with mental measurements devote adequate space to these problems, yet scattered through many journals and books will be found interesting and useful material which has never been drawn together for training purposes.

Thorough chronological review of this literature might be worth while for the new practitioner. But to include many details would probably be out of line with the practical aims of this somewhat

brief manual. Therefore this discussion will center around those problems that will repeatedly recur in setting up any new test of aptitude or achievement and proving its worth. That the technical research worker in testing will need broad knowledge of existing tests and their background and of statistical tools goes without saying. Neither of these broad fields could be covered adequately in a volume twice the size of this one. Therefore the reader will need to bear in mind the limited scope of material discussed here and apply it as the limitations of his technical background make it possible to carry it effectively into practice. Research workers in test development may find the treatment of the subject rather elementary. Teachers and civil service examiners should find little difficulty in understanding most of the material, even with very little technical preparation in statistics or psychological testing. It is hoped, moreover, that they will be encouraged to broaden their background through some of the references given in this book.

Criticisms of Examination Techniques. The attacks made upon the examinations in courses often result from the ambiguous wording of questions. If a student misinterprets a question, there may be two causes of his failure to grasp the intended meaning. One is, obviously, that he has insufficient comprehension of the subject matter upon which the item is based. The other, less obvious, is the failure of the instructor to recognize that other legitimate interpretations are possible besides the specific one he had in mind while writing the question. If the second of these causes operates to lower the student's score, the item can have little validity as a test: it does not measure the attainment of a teaching objective.

A second line of attack stems from the old problem of the relative merits of essay versus objective questions, and it amounts to saying that the latter usually place too much emphasis upon rote memory and too little upon well-organized thinking. This controversy will be treated in more detail later, but at this point it may be well to clarify the educational objectives to be reached in any course. If instruction is primarily for the purpose of enabling the student to "parrot back" what the instructor or the textbook said, it will then be relatively simple to convert verbatim statements from lecture or text into completion items to measure accurately the degree of this accomplishment. True-false items or multiple-choice items

may follow the exact wording of the instructor or book, thus calling for mere recognition of a relatively meaningless lingo in its original form as distinguished from other equally meaningless verbiage which is slightly modified. Not every student will use rote memory in answering such questions. It is possible to reason out many of them, but if the instructor does not know how many of his students are reasoning and how many are rote memorizing, even among those with high scores, how can he tell what teaching objective he is accomplishing?

Not all objective questions are like those described above. The reader will find examples in later chapters to prove that clear thinking and application of knowledge may be the only satisfactory way of arriving at the correct answer to objective questions properly designed to call forth this level of response. Teachers probably seldom stop to think what is the simplest mental process that will enable a student to figure out the correct answer to a question. Untrained in the recognition of clues to the correct answer that have nothing to do with content, but rather with the form of the question, some instructors naïvely go on assuming that knowledge of the subject matter is always essential for success. Students soon become skilled in the detection of such clues, if they are frequently encountered in examinations. Teachers therefore need a training manual that will help them cultivate the habit of determining, so far as possible, both the minimum amount of information or mental skill necessary to answer each question and the mental processes that must be carried through in arriving at the right answer.

A third angle from which some examinations can be criticized is the lack of objectivity in grading. This comment applies, of course, to the essay type of test. Granting that some professors are too lazy or indifferent to strive for objectivity, and that others are prejudiced in favor of brunettes, the reader can pass on to the more conscientious scholar of faculty status who is unaware of the prejudices and other irrelevant factors which affect his grading. His best efforts at questions are often too vague to bring any unified interpretation, and insurmountable problems of grading are the inevitable result. Some help can probably be given in enabling him to reach a reasonable standard of objectivity in essay tests if he is willing to train himself to achieve it. He may also learn when the use of essay tests will

best serve his purpose and what teaching goals may better be evaluated by means of various other types of objective tests or items.

In the recent past, many of our best universities have apparently chosen their staff members on the basis of their contributions to research or their writings in their specialties rather than their ability to teach. Interest in selling the subject to others and capacity to explain matters clearly do not necessarily accompany high achievement in research, and even writers of wide reputation may not teach effectively. Campaigns to improve college teaching have from time to time been started. Some have perhaps partially achieved their objective, but no such movement could be considered successful that did not attempt to discuss problems of test construction. Promotion of interest in development of skills to measure progress is the first step in this area. This should be followed by distribution of information from research studies with tests and finally by practice in fundamental principles applied to the subject-matter area of the specialist. If this were carefully done, there would be fewer complaints on the part of students to the effect that grades do not mean anything anyway. Instead of dispensing with grades altogether or minimizing their importance, teachers should strive to make examining evaluations more objective and thus more meaningful. They can make examinations worth while instead of a farce, more comprehensive instead of more superficial, and a means to an end instead of an end in themselves.

Goals and Objectives. Critics of our educational system sometimes maintain that grades rather than mastery of the subject matter become the goal of the student, especially if the situation is highly competitive. To minimize this competitive emphasis, they would eliminate quizzes and examinations. Yet the serious student hardly feels satisfied with a course in which individual evaluation of knowledge, skills, and abilities is either strictly subjective or entirely absent. He wants to know where he stands and also that his standing is based upon something more than mere personal opinion, perhaps hastily formed on fragmentary evidence. He wants his progress measured.

There is probably some truth in the argument that grades frequently become the student's objective and that, once this objective is attained, the subject matter is forgotten. Too often the grade

represents not a comprehensive, useful knowledge of subject matter, but a few lucky guesses at what the professor meant by his questions or perhaps a clever detection of clues to the right answer, which the instructor never realized were in the examination. In some cases it might represent an exceptional rote memory for verbatim statements, the thought of which could never be applied in any practical way since the central idea was never learned. If measurement of progress were valid, the grade would mean achievement of definite goals within the subject matter itself or mastery of skills that are important in the field. Thus the real teaching objectives themselves would become the goals without which no passing grade would be probable.

To state that all academic achievement can be evaluated with complete objectivity would be placing too much confidence in present-day test techniques. Perhaps complete objectivity is not an ideal which educators would ever want to attain. Some phases of scholarship show themselves in ways that are not subject to measurement with formal testing techniques. The suggestions made here apply to those phases not now adequately measured in most situations because many teachers, well prepared in their fields, have never given much careful attention to problems of measurement.

In the field of personnel selection, it has become increasingly evident that new tests are often needed that more closely fit the specific job. Standardized tests have been developed which select very well for some classes of positions. But often an industry has a unique class of positions for which no adequate test exists. To state that not every standardized test does what is claimed for it in its manual of directions seems like a needless elaboration of the obvious. Some of them are probably not worth the paper on which they are printed. All of them need careful, critical review in terms of their usefulness in the specific situation. After all these considerations have been thoroughly taken into account, the personnel technician may find that he will have to start from scratch and devise a new test of his own to fit his particular problem. If this is his conclusion, he will need (1) facility in converting material into good test items and (2) skill in appraising each item after a tryout of his preliminary form of the test.

Public personnel agencies have cooperated very well, on the

whole, in test-development projects. They have done this largely through the Test Exchange Service of the Civil Service Assembly of the United States and Canada, located in Chicago. The writer has had considerable contact with the material in various fields lent by this agency, and he has observed considerable variability in the quality of tests furnished by different agencies over the country. Although a few of the agencies furnished extensive research data with many of their tests which were of excellent quality, the majority of the agencies furnished tests full of instances in which basic principles had been disregarded. These examples indicated clearly the need for better-trained examiners.

As our culture rapidly changes from year to year, tests change in difficulty. Obsolete items become more difficult, and perhaps vocabulary changes in difficulty as certain words are used with greater frequency in newspapers and magazines as well as over the radio. A few examples may make the above generalization clear. On the Revised Stanford-Binet intelligence test, Form M, there is a telephone of a shape readily recognizable to the child who was tested at preschool age during the years immediately following the standardization of the revision in 1937. Today, however, comparatively few such telephones are in operation, and newer models bear only limited resemblance to the older one in the perception of preschool children, particularly in larger cities. Recognition of this test object has become a more difficult item than it was when the test was constructed. In localities where Diesel locomotives are rapidly replacing steam locomotives, the old-fashioned engine included in this test is a strange object to young children. A currently printed mechanical-aptitude test standardized a number of years ago is said to contain a Model T Ford part, not recognized by most younger persons familiar with present-day automotive equipment. The word "synchronize" became much easier for many people after the introduction of the sound film, concerning which the advertising matter made frequent use of the expression "synchronized with sound effects" and other phrases that made its meaning clear from the context. On the Wechsler-Bellevue Intelligence Scale is a picture which is supposed to be incomplete because the man's tie is missing. Changes in men's styles introducing a new type of sports shirt seem to make this item more difficult than it was at the time of standardiza-

tion, since the picture looks quite conventional now without the tie.

All the above illustrations, as well as many more that could be given, indicate the need for new revisions of existing tests, and, in some cases, new tests to replace older ones, now inadequate. Norms established a number of years ago may no longer be correct in a rapidly changing culture like ours. Difficulty of items, assuming familiarity with certain objects or practices common in the environment, may be changing to an unknown degree. Hence it is important that research in test development and test evaluation be in progress constantly. Workers in this field who can qualify will be likely to find employment readily in branches of the armed forces, industries, governmental agencies, and large universities. Skill in test construction is often an aspect of their training which is almost completely lacking when they first go on the job.

Whether this book finds its place as a text in education or psychology courses, as collateral reading to these courses, or as a training manual for personnel workers on the job in government or industry, it should have practical applications such as those described above. Its coverage is limited to often neglected aspects of testing, and it is not to be considered in any sense an adequate course of preparation for professional testing work.

Classification of Tests. Among the most thorough breakdowns of mental measurements will be found the classification given by Greene.¹ An outline in his table of contents includes other material on statistics and related problems, but it is as complete as any survey of existing tests up to the time of its printing. Super² classifies tests used for vocational guidance from a somewhat different viewpoint. Some of his main divisions correspond almost exactly with Greene's, but others are presented according to a different scheme. Both authors review critically a wide variety of commonly used tests. In other books on mental testing, the scope of material seems usually to be more limited, but a variety of approaches to the problem is possible. Here is presented an attempt to combine the best features of several outlines without giving details or existing examples. A

¹ Greene, Edward B., *Measurements of Human Behavior*. New York: The Odyssey Press, Inc., 1941.

² Super, Donald E., *Appraising Vocational Fitness*. New York: Harper & Brothers, 1949.

number of the categories overlap, such as aptitude tests and achievement tests, and each of the categories broken down by what is measured could also be subdivided according to whether the material was designed for individual or for group administration. Therefore, a large number of combinations of terms could be made to describe exactly any test with which the examiner is concerned. For example, one could use an individual, oral speed test of arithmetical reasoning or a group, written power test of arithmetical reasoning. To call either of these simply an "arithmetic test" would not show whether rapid computation or thinking ability was the primary factor measured or whether large groups could be tested simultaneously.

The outline given here is intended to show the less experienced reader how broad the testing field of today really is and also where the limited scope of this volume fits into the total picture. The clinical or industrial psychologist may have already become accustomed to different terminology for the same concepts listed in this outline, or he may prefer to list as important some subheadings that have been omitted. However, the main headings below should help define the usefulness of any new test being constructed or evaluated for a given situation.

- I. Classification according to content
 - A. Essay or free-answer items
 - B. Objective or limited-response items
 - 1. Alternative response
 - a. True-false
 - b. Yes-no
 - c. Right-wrong
 - 2. Multiple-choice
 - 3. Matching
 - 4. Completion
 - 5. Rare forms
 - C. Performance or nonverbal items
 - 1. General
 - 2. Work sample material for specific jobs
- II. Classification according to administration
 - A. Individual (to one person at a time)
 - B. Group

- C. Oral presentation of instructions and items
- D. Written presentation and responses
- E. Speed tests (presented with a short time limit to measure the amount correctly done in a short time)
- F. Power tests (with a liberal time limit in which all can finish)
- III. Classification according to scoring
 - A. Tests requiring qualitative evaluation by experts
 - B. Stencil-scored forms
 - C. Machine-scored forms
- IV. Classification according to application of results
 - A. Informal, teacher-made tests
 - B. Standardized tests
- V. Classification according to what is measured
 - A. Intelligence
 - 1. Verbal
 - 2. Nonverbal
 - B. Specific aptitudes
 - 1. Artistic
 - 2. Clerical
 - 3. Linguistic
 - 4. Mechanical
 - 5. Musical
 - 6. Scientific
 - C. Achievement
 - 1. General (all school subjects)
 - 2. Diagnostic for each fundamental subject separately
 - 3. Job selection for a specific position or class of position
 - D. Personality and adjustment
 - 1. Self-inventories, covering one or several isolated traits or dimensions
 - 2. Rating scales
 - 3. Interview procedures
 - 4. Projective techniques (ink blots or pictures)
 - E. Interests
 - 1. Vocational
 - 2. Masculine-feminine
 - 3. Recreational

F. Attitudes and values

1. Social
2. Racial
3. Political
4. Moral
5. Religious

It will be important for an examiner thinking in terms of the testing needs of a particular teaching, vocational-counseling, personnel-selection, or research situation to fit the proposed measures into this framework or at least into one similar to it. The above step would be equally necessary in his thinking, regardless of whether (1) he finds a standardized test or series of tests already available that seems to fit his purpose or (2) he finds it necessary to construct tests appropriate to the situation.

Definition of Concepts Considered. To make clear the general principles and examples to follow, the discussion must include at this point some attempt to define terms to be used. *A test from the viewpoint of psychology and education might be defined as an organized succession of stimuli designed to measure quantitatively or to evaluate qualitatively some mental process, trait, or characteristic.* This is a very broad but reasonably accurate and usable definition. The stimuli may take the form of written or oral questions, number series, geometric figures, musical tones, pictures, forms, etc. All these stimuli are essentially forms of energy that act upon the organism, producing a response. They need not be entirely sensory stimuli. They must be in some kind of a planned sequence or meaningful pattern in order to arouse any response that can be measured. Another way of stating the same requirement is that the items must be interrelated in some way or in keeping with the purpose of the whole.

The purpose of a test must always be to measure or evaluate something. *It must yield some kind of score or at least a descriptive category, perhaps both,* such as an "IQ of 84," "dull," "normal," "a percentile of 95," or "very superior." The score is of no value by itself. It must predict something, such as how well a man will be likely to succeed on a job, or it must compare him with other individuals

in some particular respect, such as dominance-submission, and show where he stands in a group in that trait.

Finally, what can it measure? The answer theoretically might be any mental process of which the human mind is capable or any distinguishing mental characteristic of a person. However, there are many such traits and processes that are not yet measurable, and there are others which are perhaps very difficult to isolate from uncontrolled factors or variables. Our present tools of measurement are crude and sometimes perhaps quite inadequate. It is generally recognized, for instance, that we do not measure intelligence directly, but rather indirectly through performance with materials that cannot be completely divorced from cultural factors. If an individual deviates too much from the culture in which the test originated, the test is unfair to him. African natives could perhaps originate a measure of intelligence which would put Americans at a distinct disadvantage. Practically, then, *a test must be proved to measure what it was intended to measure before it can be used with confidence*. No device or series of devices has yet been definitely established to predict teaching success. Numerous areas can be mentioned in which much more research is needed. In conclusion, then, a test may be said to be *intended* for a defined purpose, whether it is far enough developed to be used with confidence for that purpose or not. For most of the qualities mentioned in the outline in this chapter, at least some degree of success in measurement has been achieved.

As broadly defined in this discussion, a test is not very different from an experiment in psychology and education. In most instances, even though many of the same requirements apply to both, there is a distinction in aim. Experiments are planned with the major interest centered around a theoretical problem. An experiment is an attempt to understand a process as such, not to help an individual in his total adjustment. A test, on the other hand, is not pure science, but applied science. It is not theoretically oriented, at least not primarily, but is designed to improve the *individual* in some phase of his adjustment. It may diagnose his reading disability in an effort to locate his particular weaknesses in that area. An understanding of his strong and weak points will then enable him to practice in a way that may improve his reading. Or the test may help to

place him in the right job, one in which he will be genuinely interested.

The terms *aptitude*, *capacity*, and *talent* could perhaps be used interchangeably. Although some attempts have been made to distinguish among them, no differences seem to be generally accepted. If *aptitude* can be taken to mean a characteristic that will enable a person to learn a group of skills or field of knowledge rapidly to a point of efficient application, it could then roughly be distinguished from *achievement*, which would designate the amount of such learning actually completed. In other words, aptitude would be basic and would be theoretically present before opportunity for learning had presented itself. It would represent the hereditary factor, which would determine the limitations of learning in a particular field, regardless of environmental factors that would encourage learning to a maximum degree. *Intelligence*, or *general intelligence*, as it is sometimes called, is nothing more than an over-all aptitude. It is the facility with which one can learn or solve problems dealing with a wide variety of materials. Although there are a few exceptional cases in which marked improvement in intelligence-test scores has resulted from improved environment, it is likely that the better scores are the result of improved motivation rather than of any change in the basic intellectual level itself. A culturally poor home restrains the exercise of intelligence just as a splint inhibits a part of the body temporarily so that it cannot function in its normal fashion. An emotionally sick person from an unfortunate environment may perform far below his normal level until his mental conflicts are resolved. Then he will come up to whatever his limit is—average, superior, or genius. By and large, however, an imbecile cannot be made average, however skillfully he is taught, and an average person cannot be educated into a very superior individual.

Students have often asked the writer for an "aptitude test" with no specific designation whether one in science or clerical work was desired. Further questioning often revealed that a test of vocational interests or a test of intelligence was wanted. This shows the popular lack of discrimination with regard to tests and the descriptive terms for them. The reader can best designate an *aptitude test* as a test of more limited scope than an *intelligence test*. The discussions

to follow will treat the former in considerable detail and occasionally mention the latter, which is not the primary concern of this volume. The reader will need to keep the distinction in mind.

In the practical situations of measurement, a sharp line between measures of aptitude and measures of achievement is difficult, if not impossible, to draw. Mechanical-ability tests can be found that are designated as measures of "aptitude" which can be answered better by individuals who have grown up in an environment where tools and machines were frequently present to observe. Although native factors would affect what was comprehended, these tests could not be said to be free from the achievement factor. They measure both aptitude and achievement at the same time, and so do a large number of others.

When a personnel office or agency wants to select the best candidates for an entrance-level job (one not requiring specialized previous experience), the employer realizes that he must start from scratch and train new men on the job itself. He wants men who can be trained rapidly and who will be successful enough to stay with the company or agency. He does not expect specialized knowledge or skill at the time of employment. An achievement test is not called for here, but rather one of aptitude. For promotions to higher, more responsible positions, aptitude, acquired knowledge, and learned skill are needed. The emphasis in examining would here be upon the amount already learned. The same would apply to situations in which new appointments were made from outside the agency or industry instead of promoting old employees upward.

Teachers are concerned with the two concepts discussed here from the standpoint both of vocational counseling and of examining in courses. Final examinations, if they are good ones, usually represent some degree of aptitude and of achievement, measured in a combination appropriate to the objectives of the course. If an undergraduate, introductory college course in a science did not give the individual some idea as to whether he had aptitude in this line as well as how much he had achieved in one quarter or semester, there would be something missing in the quizzes and final examination which should be further scrutinized. Perhaps the questions were too superficial, or possibly they were too consistently verbatim from the textbook, calling primarily for mere rote memory.

There are a few other concepts that merit attention at this point. One of them has been discussed already: *whether a test measures what it was intended to measure is called the validity of the test*. There will be more to say about this concept later, but a brief example might make it clear. If, through ignorance, one were to construct a barometer with the intention of measuring atmospheric temperature, the instrument would not be valid, since it measures atmospheric pressure instead of temperature. An instrument may be carefully constructed and accurate or consistent in what it measures, but it may not do what it was meant to do. This statement may apply to a psychometric procedure developed by a highly skilled psychologist or educator. Validity cannot be safely assumed without proof. *Reliability* is high if a test gives consistent results. If chance plays too great a part in determining the score, a psychometric procedure is unreliable. The validity of a test refers to its accuracy with reference to the particular trait it is intended to measure, whereas reliability refers to the consistency with which it does so.

A *subtest* is a division of a test. There may be three to ten or more subtests in a test, and each of these major divisions usually consists of several items. An *item* is best defined as a single task or question that usually cannot be broken down into any smaller units. A *battery* of tests usually means a series of measures given to one person. These need not necessarily have been prepared by the same author nor standardized upon the same group so long as the series taken as a whole has some meaning. The results of the various parts of a battery are usually plotted on a *profile*, or graph, on the same scale, such as percentiles, which show at a glance the individual's strengths and weaknesses.

Other technical concepts, which can best be clarified in a more lengthy discussion, will occur later. Those defined above seem basic to an understanding of what is to follow, and from this broad survey of the testing field the reader is ready to progress into a consideration of the first step in test construction.

PLANNING THE EXAMINATION AS A WHOLE

IN MANY situations, it is essential to measure progress in learning to determine the extent to which important objectives have been reached by the individual or group. In many other settings, it becomes necessary to attempt to predict future attainment. As Adkins ¹ and associates state this point:

To be useful, a psychological test must provide inferences about people, just as chemical and physical tests provide inferences about things. . . . A properly constructed psychological test can provide valuable information when it is used in an appropriate situation; if a group of items are simply thrown together, however, the result may look like a test, but probably will be useless.

In developing her point further regarding planning, Adkins stresses the necessity of discriminating between those persons who are desirable and those who are undesirable in a given respect. Unless this can be done, no screening can take place either in our educational system or in employee selection. Vocational counseling would be practically worthless without some valid measurement. Whether we like it or not, our world is competitive. If competition is the pattern of our American culture, those who set up the rules of the game have as their primary responsibility the difficult job of seeing that the best man wins in a manner considered fair by those in a position to know.

¹ Adkins, Dorothy C., Primoff, Ernest S., McAdoo, Harold L., *et al.*, *Construction and Analysis of Achievement Tests*. Washington: Government Printing Office, 1947.

Defining the Purpose of the Test. After determination of what is to be measured, such as skill in typing or ability to do abstract thinking, the problem arises of determining the degree of refinement of measurement called for by the given situation. If a rough screening device is needed to distinguish bare passes from failures, the test will not have to be precise enough in its commensuration to rank superior individuals exactly in order of merit. If there were very little competition for jobs as carpenters in a municipal civil service system, for example, the agency would have no alternative but to accept mediocre men to fill the vacancies so long as these men could meet a minimum standard of performance. Discrimination between excellent and very good carpenters on the examination would probably be beside the point. The length and thoroughness of the planned examination could be adjusted accordingly. If there were plenty of carpenters on the labor market, however, and competition were keen for what few jobs were open, discrimination at the top would be essential to rank these men in order of merit. This would be necessary in order that appointments from among those passing would not be made on some arbitrary basis alone. Therefore, items would have to be included which were actually difficult enough to be answered correctly only by candidates of marked excellence.

Teachers and counselors are often concerned with problems of diagnostic measurement, which is a most exacting field. Weaknesses and strong points in detail must be picked up by such diagnostic devices. Subtests must be planned each of which will be likely to have some diagnostic significance. The test constructor must have prediction in mind in planning either employee selection or diagnostic devices.

The kind of examining procedure to be used will depend to a large extent upon the content of the job or training course for which it is intended. A well-planned academic course is not a random presentation of factual material or theories; it has definite objectives. Usually these objectives are carefully mapped out ahead of time by the thoroughly prepared teacher. Ordinarily these goals go far beyond the mere memorization of facts into the application of knowledge or skill and the comprehension of many relationships among the ideas presented. Thus the examining procedure must be based

upon a careful plan for the course as a whole. This topic will be considered in detail later.

Without job analysis, employee selection through tests can have no foundation. Basic to all personnel selection by a scientific method is sound job classification. Upon a solid foundation of job analysis, testing can have a definite purpose. This topic will also be developed later.

Finally, the examination as a whole may have to consist of other parts than a written test, if it is to be valid. This is particularly true of employment tests, which may include a performance part or work sample, an interview to evaluate personality factors, a rating of training and experience, and a written part. The purpose of the written part will then depend to some degree upon its relationship to the other sections. If the interview, for example, evaluates the candidate's fluency in verbal expression, the written section need not duplicate this objective. The job analysis will probably give an indication of what factors can best be evaluated by means of the written test.

Problems of Test Administration. Perhaps the most important single rule for good test administration is that testing conditions should be as uniform as possible every time the same examination is given. In planning for this, the test constructor should provide careful instructions for answering different types of questions and for recording answers. Sample questions explained in this respect will increase validity, since no failures are likely to result from lack of understanding of what is wanted. In college or high school classes accustomed to objective tests, an examiner may argue that instructions are a needless elaboration of the obvious, which may be true if previous tests have followed much the same pattern. On the other hand, there are many groups taking some kinds of tests that contain individuals unaccustomed to multiple-choice or matching items, for example, having been out of school for a long time. It seems better to be sure to put the idea across to the slowest examinee than to neglect a bewildered few in order to avoid boredom for a moment on the part of the majority.

To secure best results, instructions to examinees should be printed on the test forms or manual, and should be read aloud by the examiner as examinees follow, or read silently by examinees only. The

problem of how much explaining to do should not be left entirely to the examiner, especially if several examiners will give the test, since these individuals may differ to a marked degree in their extemporaneous explaining ability.

Perhaps the most difficult situation to standardize is machine operation. No matter how carefully planned the situation, there are likely to be minor differences in smoothness of operation owing to mechanical condition of typewriters, bookkeeping machines, calculators, etc. Planning for a reduction of these differences to a minimum often involves specifying a given make and model of machine "in good working condition." The quoted phrase is subject to a variety of interpretations, depending upon who inspects the equipment immediately before use. Expense of new equipment is often prohibitive. If complaints occur, there is sometimes little ground for judging whether a candidate is rationalizing in an attempt to secure a chance to repeat in the hope that he will do better next time. Such problems are a definite disadvantage in the administration of performance tests.

Again, a lack of uniformity in the written-test situation arises from differences in reading comprehension and speed. When a marked deficiency in these skills is present, it may invalidate a test for an individual who can do a job well that does not require high efficiency in reading. Educational level of those taking the test must be kept in mind in planning for both the time limits to be prescribed and the complexity of sentence structure in instructions and items. Vocabulary other than technical terminology essential to the field may pose a question of administration: whether the examiner should define a word for an examinee if that word is not an essential technical term and if without it the individual cannot understand the question. Of course, it probably would have been better to avoid the difficult or unusual word in the first place, unless the question is designed partly to determine whether that word is correctly understood.

Planning the time limits for a test that is to be given for the first time is not easy. If speed is important, the examiner will want to allow a time short enough so that most or all individuals taking the test will be unable to finish before the stop signal is given. This principle would be followed in setting up material for a plain copy-

typing exercise or a series of rapid-computation problems. The difficulty of the material in either of the above cases would remain fairly constant throughout, and it would be easy enough for any well-prepared examinee to do almost perfectly if allowed enough time. Such a plan would constitute what is termed a *speed test*. Its greatest usefulness lies where skills are to be evaluated, such as reading, computation, machine operation, proofreading, and shorthand.

When the situation calls for measurement of factual knowledge and its applications, the plan is usually different. Time limits are liberally planned so that everyone who works at a minimum acceptable speed can finish all of the items. The writer has found that it is impossible to lay down any arbitrary rule for the time allowance on various kinds of objective questions. Although suggestions of four true-false questions per minute and two multiple-choice questions per minute have been made, it is doubtful whether these can always be defended. Among factors that govern the time taken to answer questions are the habitual reading speed of the individual, which bears some relationship to his educational level, the average length of the questions, and the kind of mental process required to arrive at the answer. Automotive mechanics or junior high school pupils could not as groups be expected to read as many words per minute as social workers or law students. Items calling for association of historical events with names are likely to take much less time than those calling for the best in a series of four administrative policies in an industry or governmental agency, even when both tests are in four-choice, multiple-choice form. The former will probably contain about half as many words as the latter. After reading the item, the examinee makes the decision as to the correct or best answer; this is a matter of simple recognition in the event-name questions, but usually a process of complex thinking in the administrative items.

The test constructor can learn only through experience to make fairly good estimates of the time limits to set, and often he will not be sure that these are appropriate until after the test is actually tried out on a group. Sometimes he may wish to combine the two plans described here to reward the quick and accurate thinker. Whatever his plan, he should reveal it in his instructions if those taking the test

are to regard it as fair. Some comment such as, "You will have plenty of time to finish if you do not waste time," or, "Work as fast as you can and still be accurate; you will probably not have time to finish all the questions in the time allowed," will give the individual a *mental set* for power or for speed. He will doubtless appreciate knowing what to expect in the way of a time limit, and, if it is short, a statement of the exact number of minutes will, in most cases, be more reassuring than alarming.

Closely allied to the problem of time allowance is that of length. In planning the length of a test, the test constructor finds necessary the evaluation of several factors, some of which have to do with administration, before he can make an intelligent decision. In a comprehensive examination for a graduate degree or in a competitive examination for a very responsible technical or administrative position, the subject-matter area is broad. It must be covered thoroughly and diagnostically. A six- or even eight-hour ordeal is not uncommon for this kind of purpose, and perhaps there is some justification here for unusual length. Allowing a "coffee" or "smoke" at an appropriate division point will doubtless increase efficiency and reduce neurotic symptoms somewhat. Other situations, however, make briefer coverage of subject matter desirable in terms of expense of administration, etc.

The question of reliability when length is decreased enters into such a decision, and a later chapter will treat this in more detail. Here it can be sufficiently covered by stating that lengthening a test will increase its reliability *provided* that the added material is of equal or better quality compared with the original material. The device of adding poor items for the sake of greater length is a waste of time.

Where a large volume of examining must be done at minimum cost, there are a number of problems of administration that have to do with the material itself. Answer sheets not only provide for easier scoring without time lost leafing over pages of a booklet, but they permit reuse of the booklet many times. If a 10-page booklet is marked up by 2,000 people so that the entire copy has to be filed and cannot be reused, this is a wasteful procedure. The same questions could be answered on one separate answer sheet, reducing paper consumption to one-tenth the volume.

The greatest problems of cost are encountered when interview procedures are applied on a large scale as a part of the examining process. If the many man-hours required for this interviewing can be replaced, even partially, by written items measuring the same behavior characteristics, this compromise solution will often be necessary. Since personality testing is outside the scope of this discussion, no further comment will be made on this point. A similar problem of prohibitive cost may occur, however, in performance tests requiring that a candidate for a position consume a considerable amount of valuable material in demonstrating what he can do. To have a carpenter actually cut good lumber into pieces that could be of no further use would be a costly examination that would only serve to increase the housing shortage. The foregoing is not to be considered an argument against performance tests, but merely a clear example of what might result from poor planning concerning administrative costs.

Cheating on examinations is a social phenomenon that has many angles, most of which have nothing to do with the structure of the material itself. Only a dynamic understanding of motivation, attitudes toward authority figures, and group morale could lead to possibilities of complete elimination of this widespread practice. However, many examiners believe that planning materials so as to reduce likelihood of cheating to a minimum is one of their obligations. If seating distribution must be close because of the limited size of the room, two equivalent forms of objective tests distributed to alternate seats may help. Three forms are not uncommon. The same items may be used in scrambled order, with choices scrambled on multiple-choice questions. This is not a sure preventive measure, but it makes copying difficult, while otherwise an item number, a T or an F, or A, B, C, or D can easily be grasped at a brief glance. Many situations will not call for such precautions, but numerous others will.

The Scoring Problem. Scoring is most easily done if answers are lined up in a straight row. They can be either at the left or at the right of the questions. Still better, an answer sheet can be provided separately. Uniform markings on all papers save time. In the writer's experience, circling T or F results in the most rapid scoring of true-false items. Several clerks in a public personnel agency were tested

for accuracy of true-false scoring with various systems of marking, and this method proved fastest and most accurate. For multiple-choice items, the underlining of the correct response takes more time in both administration and scoring than simply indicating the correct response by writing the number 1, 2, 3, or 4 or the letter A, B, C, or D. For matching items, greater speed and accuracy of scoring results when the answers are written down by letter or number than when lines are drawn from the list on the left to the correct member of each pair on the right. Large-scale scoring by five clerks checked by the writer seems to establish fairly well the superiority of the methods described above for matching and multiple-choice material.

A minor problem arises when, to save paper, the printer or duplicator insists upon single-spacing the answer sheet in typing the stencils. Very close spacing of answers to the left or right of questions crowded together on the page also decreases both speed and accuracy of scoring, according to the writer's experimental data. Double-spacing of answers increased speed of clerks by approximately 10 per cent, while their accuracy went up an average of roughly 5 per cent. Only one of the five clerks was unaffected in speed and accuracy by the change in spacing. The writer found his own speed 10 per cent better and his accuracy 10 per cent better with wide spacing. Although there is a possibility that chance affected the change in number of errors, there seems little doubt that speed was really improved.

Of course, machine scoring is the most economical procedure in very large volumes of scoring. High rental costs on International Business Machines Corporation (I.B.M.) scoring equipment would prohibit this procedure where very limited amounts of material are to be processed, but clerical help rapidly becomes more costly than the machine when scoring jobs run into thousands of papers per week. An increasing number of standardized tests are now becoming available in machine-scored form. The process may be described briefly in several stages.

First, an answer sheet is provided that has 200 answer spaces on each side. A set of such spaces would look like the sample on page 24. Space A has been blackened to show the method of indicating the first of five choices in a multiple-choice question. The same

	T	F			
	A	B	C	D	E
1.	:::	:::	:::	:::	:::

method could be used to show that the answer to true-false question 1 is *true*. Short matching items could be set up for this sheet.

Secondly, the test must be taken with a special I.B.M. pencil, which makes marks that will be picked up by the machine more readily than ordinary pencil marks. If more than one space is blackened, inaccuracy in scoring results, and papers have to be inspected for this and for other markings in the wrong place on the page, but scoring achieves very good accuracy if instructions are followed with regard to using the answer sheet and if the machine is very carefully checked for adjustment frequently during scoring.

Next, the sheets are warmed and dried in a part of the machine. Meanwhile, a stencil card has been cut for right answers and one also for wrongs. Holes expose the marks for the right answers on a prepared key answer sheet. The stencil is inserted and the key checked to see if the machine reads a perfect score. Adjustments are made. The machine can be set to give a score equal to rights minus wrongs or number right.

Papers are then inserted one at a time, and the scores are read immediately from a dial. One person can easily run five or six papers a minute through the machine and record scores. Two persons can do it even faster. The preliminary adjustments are likely to take much more time than the scoring of smaller groups.

Weighting Scores on Composite Tests. Much controversy has arisen over the weighting of scores on parts of an examination. Sections are often weighted according to the test writer's subjective judgment of their relative worth. A better plan is to make a distribution of subtest scores for each section after tryout and then give greatest weight to the subtest yielding the greatest dispersion (standard deviation). When this is done, there is presumably greater accuracy of measurement, since the dispersion for the test as a whole is thus made wider. Correlations of weighted scores with unweighted scores usually turn out so high, however, that some examiners doubt whether the slight gain in reliability is worth the amount of increase in complexity of scoring. The more operations involved in scoring,

the greater are the chances for error. Time may be an important factor in some scoring situations, and the simplest plan will turn the work out quickly. Hence, from a practical standpoint, weights can usually be left out of consideration in planning a written test. If an interview, a performance section, or a rating of training and experience are to be incorporated into a personnel-selection procedure, the weight of the written test in relation to these will have to be considered in planning the entire procedure. Ordinarily the written examination gives the widest dispersion, though not always, and if it measures what seem to be the most important aspects of the performance of the job, it probably deserves greater weight than other divisions of the process.

Corrections for "guessing" right answers constitute another subject over which there has been much controversy. One argument for correction of scores for chance is that guessing should be discouraged in order that the score can represent real attainment unaffected by "lucky" guesses. For multiple-choice items with four choices, there would be theoretically only one chance out of four of answering one right without any knowledge whatever of the subject matter. Actually, however, the guessing factor is probably not usually mathematically predictable by the laws of chance alone, since the typical procedure of the examinee who does not know or is not quite certain of himself is a negative one. By common sense he figures out that the correct answer could not be D. He knows just barely enough to recognize B as wrong, and he has no basis other than chance for deciding whether C or A is correct. Therefore he has one chance in two of answering the question right. If this reasoning is really typical, as many students report that it is, chance is an uncontrolled variable in the situation, and its influence is somewhat unpredictable.

In true-false tests, the right-minus-wrong procedure is quite common in correcting for chance. If it is used, most authorities agree that those taking the test should be told how it will be scored. The penalty for guessing will not discourage it in every case, since some people are more willing to take chances than others, but most test constructors explain this scoring procedure in the instructions. The correction is particularly applicable to speed tests. For example, suppose that subject A is a very careful worker. He covers 15 out of 50 items in the time allowed, but answers them all correctly. Sub-

ject B, on the other hand, is fast and careless. He likes to take chances and thinks that the more items he covers the more likely he is to be "lucky" and make a high score. B comes out with 30 attempted, 15 right and 15 wrong. The case is hypothetical for the sake of simplicity, but it is not an exaggeration of actual extreme cases in the writer's experience.

The question for the test constructor is how to score these individuals. To give them both 15, the number right, does not make sense. Rights minus wrongs would give B zero and A the same 15. These persons are a long way from equal. Assuming that the test measures speed and accuracy, and that, as is usually the case, both factors are of equal importance, the examiner could probably conclude that the accuracy score of B was chance or zero, and that the speed was worth nothing without accuracy.

However, there is a more serious question of what to do with the subject who hits a happy medium between extremes A and B. Fast workers who make a few minor errors may be more desirable than overmeticulous ones who make none but go too slowly. The degree of accuracy required on a certain job will determine the best answer to the problem.

Analysis of Content of Course for Tests of Achievement. The teacher who makes a final examination by leafing over the pages of his textbook and notebook asking himself, "What can I find here that will make a good question?" will not know whether his final product is valid or not. If he has definite teaching objectives outside of rote memory for verbatim passages, he will need to ask himself quite a number of other questions before he prepares any for his students. Here are some of them. "What do I want the student to be able to do with these facts? What will he need to do with them in later years on his job or elsewhere? What relationships should he have grasped among concepts presented in the course? What related ideas should the student be able to contribute from his own experience? What skills should he have learned?"

Then the relative importance of various topics and details in the course content will have to be evaluated in order to achieve adequate coverage in the time that can be allowed. Consideration of how many questions to ask on each topic will make the examina-

tion weight itself properly for the emphasis each idea received in the course. Thus no weighted scores will be necessary.

Ultimately the instructor must decide which form of test item will fit his purpose best—essay, completion, matching, etc. Tentatively he may map out the examination in main topics, indicating for each the type of question he wants and the probable number of them. This preliminary outline need not be followed slavishly. In fact, it may be changed quite radically in some ways after the instructor has begun to dig for material. But it is always helpful to start with a good plan rather than to proceed hit or miss.

Job Analysis as a Preliminary to Job Testing. In the field of personnel selection, time and motion studies were the earliest methods used in describing the behavior of workers in various skilled occupations in an effort to understand the abilities required. Among these early studies may be mentioned the observations by the Gilbreths,² which were rather limited in scope, leaving out of consideration almost altogether such factors as working environment, training and formal education, and level of responsibility. These early attempts at job analysis, however, led to more extensive research that proved useful in personnel selection.

One of these later developments was the job psychographic method introduced by Viteles.³ This method begins with a description of the occupation in terms of the duties performed, the nature and conditions of the work, the training involved, the related jobs from which workers may be recruited and to which they may be promoted, the advantages and disadvantages of the work, and the personal, physical, educational, temperamental, and experience requirements of the job. The examiner gathers the data by watching the work being done and by interviewing workers and supervisors. Then it continues with a rating on a five-point scale of a standard list of 32 abilities, such as coordination and visual discrimination. These ratings are plotted on a profile, or psychograph. From the psychograph, the examiner can then plan, compile, or construct a battery

² Gilbreth, F. B., and Gilbreth, L. M., *Applied Motion Study*. New York: Sturgis and Walton, 1917.

³ Viteles, M. S., "Job Specifications and Diagnostic Tests of Job Competency." *Psychol. Clin.*, 1922, 14:83-105; *Industrial Psychology*. New York: W. W. Norton & Company, 1932.

of tests to measure the most important abilities required for the job.

More recent is the plan developed by Shartle ⁴ and used by the United States Employment Service. The factors listed go into considerable detail.

Job specifications as written today may sometimes cover broad classes of positions, such as *Clerk*. Positions in this class may include file clerks, clerks who check information on forms to see that it is accurate in every respect, and other clerks who record largely numerical information. The specification gives samples of work performed in the various positions in this class and lists knowledges, skills, and abilities required at time of appointment. It gives required or desirable training and experience. All jobs within the class are assumed to have in common some general types of duties (simple, routine, clerical) and certain personal, physical, mental, and educational requirements. The assumption is probably usually correct, but there are certain stereotyped phrases in the jargon of job classification which the examiner finds difficult to define. What is the difference, for example, between "some knowledge of office procedures" and "good working knowledge of office procedures"? How difficult should the questions on this topic be to fit the descriptive terms in the two specifications? The examiner will have to interview workers and supervisors himself. He cannot depend entirely upon written specifications.

Other specifications may be written around narrow class concepts or even single-position classes. Even here the vagueness of terminology is often evident. Super ⁵ explains this problem in the statement:

The analysis of the job provides the test constructor with a list of aptitudes and traits which are deemed important in that job. But this list is subject to two serious limitations. These are the subjectivity of the evidence and the uncertainty that a particular factor, even if it proves to be important, will differentiate this job from others.

Super continues with an example of a trait that is common to many jobs—getting along with others. This characteristic may be placed in a specification because some supervisor or some job analyst ex-

⁴ Shartle, C. L., *Occupational Information*. New York: Prentice-Hall, Inc., 1946.

⁵ Super, Donald E., *Appraising Vocational Fitness*. New York: Harper & Brothers, 1949, p. 31.

pressed the *opinion* that it is important. Probably there is no objective evidence to support his view. In other words, the evidence is all subjective in the case mentioned here, and perhaps in a number of others, especially as to the degree of such a trait that represents a minimum for success on the job.

There is no known measure of the *amount* of "tact and courtesy in handling clients" required of a welfare fieldworker. "The ability to supervise others," taken from a specification for office manager, is not entirely textbook knowledge of fundamental principles and accepted practices. It is best described, at our present stage of understanding, as partly a qualitative personality function, not directly measurable. "Ability to interpret the public welfare laws of the state" has been required for local county welfare directors, but the interpretations made by successful welfare administrators are not always in agreement, even on major issues. The term is a broad, poorly defined one. "Good judgment" is another convenient cliché that is subjective if unexplained.

Many of the above examples are not distinguishing characteristics of one job or even of one broad class. They are, rather, common to a number of jobs, and therefore no vocational guidance or job placement can be made from attempts to predict on the basis of measurement of them. The test constructor, therefore, will find that he can plan to measure only a sample of the most important abilities required on any but the simplest jobs.

A partial answer to the objection that the knowledges, skills, and abilities listed on the specification are mere matters of subjective opinion may be found in the occupational-ability pattern.⁶ According to this technique, it is often possible to verify these opinions by giving a battery of tests to individuals who were rated as successful in specific occupations. If a representative group of successful people in each job or class of positions is tested, the average scores on each test can be plotted on a profile, or psychograph. This will then sometimes furnish concrete evidence of the amount of each *measurable* trait needed to do the job well. The assumption is made, and probably correctly, that standardized test results are better than expert opinions on this matter. When such a group profile is made on an

⁶ Dvorak, B. J., *Differential Occupational Ability Patterns*. Minneapolis: University of Minnesota Press, 1935.

adequate sample of traits, an individual can be given the same tests to discover how closely his profile fits that of the successful group.

The above procedure is not a complete answer for the test constructor, since he will probably be working with job concepts that are a little different in some ways from those on which such research has already been done. Also, the tests may fail to sample some important traits that no experts mentioned and therefore were left out of consideration. Obviously, some traits mentioned by the experts may not as yet be reliably measurable.

The research worker in personnel selection must recognize, then, that the specification is a mere starting point. If he is going to work on a test for carpenters, he must first smell the lumber. He may benefit from doing a little of the job himself to learn its technique, but if this is impossible, he must rely heavily upon the help of experts, especially the job analyst. The classification man is usually concerned chiefly with pay ranges, labor-management relations, departmental reorganizations, and the like, seldom with examining, unless its problems are brought to his attention. Cooperation between examining and classification specialists may result not only in more exact job specifications and examinations of improved validity but also in setting up a criterion against which to validate tests later.

The problem of the validity of personnel-selection procedures is one to be considered in a later chapter, but the test constructor must often think in the planning stage about how he is going to prove that his tests are worth while—that success on them correlates with job success. Unless those in a supervisory capacity are cooperative in setting up a dependable way to measure job success, the value of the tests may often be justly questioned. There are perhaps two kinds of criteria, immediate and ultimate. One test may predict well whether men can finish law training successfully but not whether they will be successful in the practice of law. Unless the immediate criterion measures at least some aspect of the ultimate criterion, the test is narrow in scope and of limited value.

Validation has been superficially passed over far too often. Now its importance is becoming real to those using tests in practical fields of education and employment. Validation should be planned while the test is being constructed, so that a tryout can be made on a

group that can be followed up later. The test constructor will find his ultimate task easier if he asks himself what criteria will be available to him for the evaluation of his measurement of each trait. If he sees little chance of proving the worth of a new idea, he might at least consider abandoning it.

Tentative Outline of the Proposed Test. After the analysis of teaching objectives or the job analysis has been completed (depending upon the type of test), the next step is to make an outline to guide the research worker in selecting or originating the proper number of items to give appropriate weight to each trait to be measured. In some situations, particularly employee selection, the worker will find helpful a *rough plan* at this stage prepared somewhat as follows. First he will go over a list of traits or teaching objectives such as those discussed in the two preceding sections to divide them into groups, or categories. One category will include all those which seem logically to be most easily measurable with a written test. Most such topics will be aptitudes or items of information. Another group will include skills best evaluated with performance tests. Still another will contain personal characteristics which in some instances cannot be measured or evaluated at all without psychiatric interviews or projective techniques, which are too expensive to be of value except as diagnostic clinical procedures. Some of these aspects of personality can be measured, however, perhaps through inventory or interview techniques. Finally, there might be a group of traits for which a detailed training and experience record would be sufficient.

An example of such a rough plan for an examination for the position of *psychometrician*, or for a training course in psychometrics, might read somewhat as in the example on page 32.

Since this discussion of test construction is limited to (1) measures of aptitude and (2) tests of achievement, the interview and past record may be omitted from further consideration. The next step will then be a further breakdown of the written and performance tests, if both are to be included. There may be several requirements of the job or aspects of the course content that cannot be measured by either of these devices. The research worker, as stated before, can examine only a sample of important traits, not the entire content.

<i>Important knowledges, skills, and abilities</i>	<i>Method of measurement</i>
1. Individual intelligence-test materials 2. Group intelligence tests 3. Statistical methods 4. Tests of aptitudes 5. Achievement tests 6. Personality and other tests	Written test
1. Establishing and maintaining good rapport 2. Quick and accurate timing 3. Skill in handling test material	Performance test
1. Alertness, quick grasp of new situations 2. Ability to get along with fellow workers, superiors, and public 3. Facility in verbal expression	Interview
1. Well-rounded cultural background	Training and experience record

There are several acceptable plans for arranging a written test. If it is a speed test, the items are likely to be of the same or approximately the same level of difficulty throughout. If it is primarily a power test, the items may be arranged in increasing order of difficulty from very easy to very hard. This order of difficulty may prevail within each section of items on each major topic. When each new topic is introduced, the easiest question on it comes first and the most difficult one last. The order of difficulty may prevail within a section containing one form of item, such as true-false, and then begin all over again with easy material in multiple-choice form. Still another common arrangement is a spiral form. If, for example, there were four general types of questions on the entire written examination—vocabulary, space perception, arithmetic reasoning, and verbal analogies—the spiral arrangement would give one item of each type first at an easy level. These four would be followed by four more—one vocabulary, one space perception, and so on—at a slightly more difficult level. The difficulty would thus increase through several levels before the end.

Generally, the spiral plan best fits measures of intelligence and simple aptitudes. Complex achievement tests, however, usually break down most logically either into subject-matter areas or possibly into

forms of questions, such as matching or essay. Numerous combinations of the above plans are possible. Perhaps the most objectionable arrangement is one which is illogical and therefore unnecessarily frustrating to anyone taking the examination. For instance, nothing would be gained, and possibly much good will would be lost, by introducing first a completion item about World War II generals, then a multiple-choice question on industrial personnel practices, and next a true-false statement concerning electricity, etc. If such divergent topics are necessary in one examination, logical grouping of items by subject matter would be best, with form of item a secondary problem that might be handled by subgroups under each topic. Haphazard arrangement of subject matter and item form breaks all meaningful and helpful mental sets, thus measuring resistance against confusion rather than actual capacity to think or achievement.

To clarify this outlining step in test-construction procedure, a sample may be given here that is broken down in rather detailed fashion by subject-matter area. Experienced workers will often put only the main headings into writing, but they will be likely to think of appropriate subheadings and details as the material develops. To write down these details may be helpful to beginners in avoiding what are later apt to become important omissions. Whether this preliminary outline is detailed or brief, it must remain flexible. Workers who consider it a pattern to be followed rigidly without the slightest deviation or modification will curb their own originality, making the final product less imaginative than it should be, without the slightest departure from the conventional. Changes for the better will occur to the capable, original writer of tests as he pulls items out of various source material. He may, in fact, end up with a radical revision of his original plan. If he does make important departures from his preliminary outline, he will not have lost the time spent in developing his tentative plan, which was only a necessary stage in his creative thinking.

The preliminary framework around which the worker builds his test will serve him best if it indicates main, and perhaps subordinate, headings, form of item or question, and suggested number of items in the block, or section. Sometimes it will indicate proposed time limits for the test as a whole or for the individual sections. At least

it may suggest possible working time likely to be required by each section or the whole. In some cases of employment tests, it will show the class titles in a series consisting of more than one class of positions and the subject-matter blocks that apply to each. Tests for any two classes in the sequence are likely to overlap considerably in content. Other notations may be made on the outline covering method of grading (number right, rights minus wrongs, etc.).

As an illustration of such a preliminary framework for a written, objective achievement test, a final examination in an academic course in psychological testing is outlined here. This outline is intended to be in tentative, not final, form, and needed revisions in it will occur to some students who examine it critically. The personnel specialist will note that this plan might very well also be suitable for some positions as psychometrician.

FINAL EXAMINATION IN PSYCHOMETRICS OR PERSONNEL-
SELECTION TEST FOR PSYCHOMETRICIAN

<i>Subject-matter area</i>	<i>Form of item</i>	<i>Number of items in block</i>
1. Definitions of terms	Matching	10
2. Validity and reliability	True-false	5
3. Individual intelligence tests	Multiple-choice	20
4. Group intelligence tests	Multiple-choice	20
5. Calculation and use of mental age, IQ, and percentiles	Multiple-choice	15
6. Aptitude and achievement tests	Matching	10
7. Personality inventories and projective methods	Multiple-choice	15
8. Other types	Multiple-choice	5
9. Scoring by machine, stencil, etc.	Multiple-choice	5
10. Statistical problems	Completion	5
11. Testing conditions	True-false	5
12. Interpretation of results and profiles	Multiple-choice	15
13. Principles of report writing	Multiple-choice	20
	Total items	150
	Total time about 3 hours	

Any of these subject-matter areas could have been broken down in more detail. For example:

<i>Individual intelligence tests</i>	<i>Items</i>
a. Stanford-Binet	8
b. Wechsler-Bellevue	7
c. Mental growth and retardation	5

It is doubtful, however, that these divisions of the topic have much value at this stage of planning. An attempt to follow rigidly the exact number of items predicted for each would be absurd.

The research worker will not always find so many good units of material to convert into questions as he expects at the planning stage. He may find that a topic which he intended to treat in multiple-choice form does not yield questions of this type readily from the available sources. Under these conditions, he may prefer to state his questions in true-false or matching form instead. The nature of the subject matter will affect his choice of form in ways that will be discussed in the next chapter. The preliminary outline is merely suggestive of the forms of items and the relative weight assigned to each major division of the test.

The writer of the example above, in developing his actual test, became concerned over one apparent weakness in his plan, namely the omission of any evaluation of the individual's skill in explaining the meaning of test results clearly. Although diagnostic clinical insight was not developed in an elementary course of this sort, nor is it expected on a job as psychometrician, limited interpretation of a few tests would not be an unrealistic requirement. It would, in fact, be a desirable teaching objective. Topic 13 was originally planned as a series of brief case illustrations requiring the choice of one out of four possible interpretations. The best of the interpretations of scores would illustrate a rather obvious principle in each question. There might be two or more questions per case.

This procedure, however, would not ensure the ability to give in writing or orally a logical, clearly expressed, well-organized statement of what the results mean. Practical testing situations demanded, in the writer's opinion, that an essay question on each of two typical test profiles be substituted for the multiple-choice form. Creative thinking leading to this conclusion might not have been stimulated until after the tryout if the preliminary outline had not been made first and constantly reevaluated as progress was made on the examination itself.

Another part of the entire process seemed logically to be a performance section. This was conceived as administration of four complex sets of instructions and demonstrations to the examiner as subject. Later the samples of material for this section were selected as planned. Some idea had to be germinated at the outlining stage of the appropriate weights for the written and performance parts. The latter seemed subjectively to be worth from one-fifth to one-fourth of the final score.

After the planning is completed, possibly in cooperation with a job analyst or expert, the test constructor is ready to consult his sources for material. If he merely copies questions from other tests, card files, etc., his end product will be a conglomeration of illogical sequences of items, with wide gaps here and overemphasis there. Such a quiz can hardly be dignified by being labeled a test or measuring instrument. Time is gained, not lost, by outlining. Mere random consultation of a wide variety of sources will very seldom, if ever, result in a balanced final product in any sense. Relying upon the space devoted to each topic in just one book is seldom a way out of the thinking and planning stage. Good content, when located, will fit easily and quickly into a logical preliminary plan.

CONVERTING MATERIAL INTO OBJECTIVE-TEST ITEMS

HAVING the tentative outline before him, the research worker in test development is now ready to obtain items from whatever source he can discover, buy, borrow, steal, or invent. Whatever his tactics may be, he will have the responsibility of evaluating what he compiles as he goes along. He cannot usually assume that this has already been done adequately for him if he compiles material from other writers instead of creating original items. Acceptable standards of item construction have not been universally followed, and critical review will nearly always be essential.

This chapter will begin with suggestions regarding sources and how to make effective use of them. A number of valuable rules for item construction have been compiled by Ruch,¹ Adkins *et al.*,² Ross,³ and others. However, the statements of these rules have often been in negative form only. The research worker needs to think in terms of what to achieve as well as what to avoid. Therefore, this chapter will attempt to clarify the most important rules, compiled from various authors, revising the original wording in some instances. Illustrations of conformity to and disregard of these fundamentals will be selected from several subject-matter fields. Essay questions, specialized kinds of objective questions in certain fields, and performance materials will be considered in later chapters.

¹ Ruch, G. M., *The Objective or New-type Examination*. Chicago: Scott, Foresman & Company, 1929.

² Adkins, Dorothy C., Primoff, Ernest S., McAdoo, Harold L., *et al.*, *Construction and Analysis of Achievement Tests*. Washington: Government Printing Office, 1947.

³ Ross, C. C., *Measurement in Today's Schools*, 2d ed. New York: Prentice-Hall, Inc., 1947.

Available Sources. For the teacher, test preparation is often a last-minute task of a somewhat routine and unpleasant nature, done in a great hurry without much planning. His usual sources are his textbook and his notebook, and quite logically they are most helpful; his method of using them is the point most likely to be at fault. Failure to use his head as a third source of material will probably result in his copying verbatim quotations from text or lecture, omitting a word or phrase here and there for completion by the student. Or he may change a "can" to a "cannot" or a "usually" to a "seldom" and ask the student to label the incorrectly quoted sentence as true or false. Sometimes, if he feels really ambitious, he may, in addition to copying the statement, try to locate in the chapter or lecture three inappropriate but plausible words or phrases to form the wrong answers to a multiple-choice form. The above procedures are not to be condemned entirely, but their use to the exclusion of others has frequently resulted in the justified criticism that most objective tests are limited in scope to the measurement of rote-memory functions.

Any experienced educator will find situations in which rote learning is indispensable. In such situations, its measurement by means of questions pulled directly from verbatim quotations would be appropriate and valid, provided, of course, that the questions were skillfully selected and constructed. Knowledge of English grammar, dates and names in history or current events, books and their authors, and legal definitions might be examples of material for which these types of questions would be appropriate.

On the other hand, the professional educator will usually try to determine whether material has meaning for his students, whether they can apply general principles to new problems, and whether they can generalize upon specific data. Measurement of these common teaching objectives requires originality, appropriate departures from the wording of the text or lecture, and sound judgment as to the simplest mental processes that will make possible the choice of the correct answer. Numerous instances in the writer's teaching experience demonstrate that minor substitutions of words for those of the original without fundamental change of the context will confuse some students completely, while others recognize the basic ideas just the same. These alterations therefore detect something funda-

mentally wrong with the original learning, assuming that mechanical reproduction of words of the author probably has little value in and of itself. The question which is not a verbatim quotation is likely to screen out the thinkers from the memorizers. The student who "reads between the lines" and relates new thoughts to material he already knows is not confused when ideas are clearly stated in any one of several ways. He has digested the content of the course.

Many textbooks contain thought questions, usually in essay form, at the end of each chapter. These may suggest objective test items. Some modern texts are accompanied by objective tests in supplementary workbooks for the student or in manuals for the teacher only. The workmanship that these tests represent varies from rather poor to excellent. Among the most comprehensive collections of such questions will be found those developed by Ruch ⁴ for use by the teacher of elementary psychology. Those in the workbook by the same author ⁵ are also questions that were well tried out before publication. Although a few of these have been found by the present writer to be ambiguous, even to some of his best students, the vast majority of them are clear, to the point, and well selected.

In courses where lectures form a significant part of the presentation, most teachers with any broad knowledge of their subject will branch out from the text into their own experience and outside readings rather than follow the book closely. The extent of this departure will vary greatly from one individual to another, but some degree of it appears to be desirable from the point of view of motivation to be present and to give close attention to what goes on in class. Lecture notes will form a useful source for test questions unless the lecture amounts merely to a process of transferring material from the teacher's notebook to that of the student without its going through the mind of either. In other words, the effectiveness of a lecture can be evaluated in terms of the extent to which it has provoked thought.

Class discussion in which students as well as the instructor have participated should occasionally be an excellent source of examination material. Even though such exchanges of views are inclined to

⁴ Ruch, F. L., *Psychology and Life*, 3d ed. Chicago: Scott, Foresman & Company, 1948.

⁵ Ruch, F. L., *Working with Psychology*. Chicago: Scott, Foresman & Company, 1948.

be controversial, some ideas of value usually emerge if the students are superior in intelligence. If the discussion has to end without establishing which of two opinions is correct, at least the understanding of the main arguments for each viewpoint can be measured, and probably should be.

Nonacademic situations may call for measurement of aspects of behavior not explained in any textbook, at least not clearly. This is where job analysis enters the picture again. We have encountered through job-analysis methods some general traits to be measured, but the specific methods of approaching them will become the problem when items are being developed. Close observation of the job by the examiner may bring to mind some creative ideas here. Sometimes the act of doing the job or part of it will show up the difficult points in learning it. In administering an intelligence test to a child, for example, a frequent cause of failure lies in the inability of an inexperienced examiner to maintain rapport with the child without becoming confused in the several steps in giving instructions and making an important omission. The experienced clinician is likely to have developed well-fixed habit patterns with which to meet this situation. He will not find these in any test manual, but will acquire them for himself if he is really interested in maintaining standard testing conditions. Doing the job is the only way of observing some fine points, which on an examination may discriminate between an excellent clinical psychologist and a mediocre one. There are aspects of many complex skills that have not yet been well enough understood to be described in books or instruction manuals. The research worker who is a keen observer has a chance for original discovery here on the basis of which highly discriminating test items may be made.

Sometimes expert workers on the job itself have been called upon to construct actual test items. A few governmental agencies have tried training expert accountants, civil engineers, nurses, and social workers in test construction. A condensed in-service training course, however, leaves wide gaps in background in test construction, administration, scoring, and interpretation. To make up for this weakness, thorough review by an expert in testing is required. While this procedure avoids errors of omission that might otherwise occur, even the very best workers on some jobs have little verbal facility and can-

not explain what they do. Indeed, many occupations do not call for such verbal facility. Therefore it is not surprising that few workers have any idea of how to shape up a test item. Occasionally the personnel technician can cooperate very closely with an expert in the subject-matter field in the production of items. The writer recalls his experience in preparing a personnel-selection test for steam-operating engineers. He knew nothing worth mentioning about the jobs, even after observing them for some time. A highly skilled boilermaker worked with the writer, but the former verbalized his occupation quite inadequately. The boilermaker drew diagrams, demonstrated, and tried repeatedly to explain important points. The writer made an attempt to shape each point into a multiple-choice or perhaps a yes-no question. He then asked the boilermaker to read over the question and criticize it. Gradually a few items emerged which were needed to supplement those from published sources.

Some skilled tradesmen were unable to be of much help in constructing items or in giving constructive criticism, even though their production on the job was regarded as exceptional. Their comments on questions attempted by the writer often went no further than something like, "They ain't gonna know what you mean by this'n." Trade school instructors proved to be of more valuable assistance, however, since they had often discovered enlightening facts about common learning difficulties of men in the trades. They were accustomed to explaining operations as well as doing them, but principles of uniform testing conditions, validity, etc., were frequently entirely new to them.

Nothing can replace originality of thinking when test items relating to jobs are to be constructed, but the uninformed must have some basis or background for new ideas. If no ideas seem to come, there are perhaps a few helpful hints that can be given. One is to ask an expert in the particular line to explain some of the common errors that are made by inexperienced workers. These faulty techniques, provided they are really inefficient and not just the product of other equally recognized but opposing schools of thought, make good wrong answers in multiple-choice items. The test constructor must make sure, however, that his right answer is really the only right one, the only defensible one.

There is danger of questions being framed from one school of

thought only, especially in professional social casework, education, psychology, sociology, and other somewhat controversial subjects. Academically this practice is perhaps legitimate, in so far as the student has ample opportunity to become acquainted with his instructor's viewpoint and should know what the "expected answers" are. However, in a civil service examination, and in other situations where the item writer's school of thought can only be guessed from the context, the designation of an author or school seems only fair. The question itself may need to specify, "According to Freud," or perhaps, "As the gestaltists would approach the problem." In order to make sure that legitimate opposing views have not been overlooked, the research worker in test development may find two subject-matter consultants better than one, especially if the two received their training and experience in widely separated localities. They will serve as a check on each other. If both of them accept an item as good, and agree upon the answer, the examiner is more sure of its validity than he would otherwise be. If they cannot agree upon the answer to an item, that one had better be left out unless it can be better defended.

For governmental personnel agencies, there is a valuable source of exchange and cooperation in the Test Exchange Service of the Civil Service Assembly of the United States and Canada, with headquarters in Chicago. Member agencies send in copies of tests for various classes of positions, which are in turn loaned out as needed by state, county, and municipal civil service and merit-system examiners elsewhere. Research done on the tests is sometimes reported, if it includes any statistical analysis. The quality of the items in such tests varies from excellent to very poor, as does the preparation of the examiners who wrote them. Many examiners in such agencies rely almost entirely upon copying in part the examinations borrowed in this manner, having little originality or technical training with which to create their own. Some test technicians select intelligently from these sources, editing the original considerably. Others are apparently incapable of evaluating the suitability of the material, and many mediocre questions or even undesirable ones have perpetuated themselves over the years until no one knows who originated them.

Standardized tests are, of course, copyrighted, and they cannot

legally be copied in part or as a whole. However, a working knowledge of many of these, together with some familiarity with the best of them in a wide variety of fields will be a valuable asset to anyone doing work in the development of new tests. General ideas about outlining, content, and form can be obtained from studying the best published tests. Such ideas can be used to advantage in constructing entirely original material.

This discussion may best be concluded with the suggestion that several sources are nearly always better than one for any test construction project. This is just as true of educational tests as it is of personnel-selection tests, either public or industrial. The same basic principles apply in all these situations in all stages described here. Only minor differences will be found. The kinds of items selected or written, however, will vary considerably, depending upon the purpose of the test, what it is supposed to measure, and the population for which it is intended. The sections to follow will deal with various forms of items applied in a variety of situations.

Alternative-response Items. Some of the simplest questions can be stated in a form in which they can be answered by *yes* or *no*. If the point to the question is a factual one, like, "Is a centimeter shorter than an inch?" there can be but two alternatives, the affirmative or the negative. There is no controversial issue involved. Only one interpretation of the question is possible. Therefore it is suitable for an alternative-response form. The subject matter can be elementary, as in the above example, or it can be highly technical, as in the question, "Is the standard error of the mean used in computing the standard error of the difference between two means?" The answer to this question is *yes*. No controversy is involved.

The yes-no form finds its greatest usefulness with elementary school children and with applicants for predominantly manual types of work requiring little education or verbal intelligence. Applicants for jobs such as janitor, watchman, and hospital attendant would be confused by the concepts of *true* and *false*. Especially those men who had their limited schooling some time ago probably would never have taken a true-false test before. In the writer's experience, the yes-no form is much more readily grasped by these individuals than is the true-false pattern. The multiple-choice form is likely to become too involved for them. Their reading skills are not equal to

the task unless the sentence structure is very elementary, and even oral presentation does not seem to help materially, since attention span for verbal material is likely to be short. The concepts *true* and *false* are abstract, and for many of these individuals would lack face validity. In other words, these concepts would seem entirely out of the field of the work for which they were applying.

Grade school children would find these yes-no items easy for the same reasons. Occasionally rather technical verbal information may fit this form well. Instructions for answering will vary with the situation, but for those unaccustomed to objective tests, the directions must be childishly simple if the test is to have any validity. Illustrations of a question answered in the affirmative and one in the negative are not at all uncommon. A satisfactory set of instructions tried by the writer is:

Each question below can be answered by *yes* or *no*. Read each one carefully. If the answer is *yes*, draw a circle around the word *yes* as shown below:

SAMPLE: Is the capital of Texas located at Austin? ☒ YES NO

If the answer is *no*, draw a circle around *no* like this:

SAMPLE: Would Texas be the smallest state on the map? YES ☒ NO

Work rapidly, but do not guess. Guessing will lower your score.

A machine-scored answer sheet would probably be undesirable for people with limited education because of the complexity of directions required.

A modification of the yes-no form is the right-wrong question. It would be written as a statement just as in true-false form, but would be classified in a simpler dichotomy, as:

Austin is the capital of Texas. RIGHT WRONG

There do not seem to be any particular advantages to this form.✓

The most commonly used variety of alternative-response pattern (true-false) has been discussed at great length in psychological literature. The research reported has brought forth rather conflicting conclusions, perhaps because of the wide variety of material used as a basis for the questions studied, and the differences in the skill with which they were constructed. Several advantages have been mentioned. One is the large sample of subject matter that can be cov-

ered per unit of working time for the examinees. Much more time would be required to cover the same facts in multiple-choice, matching, or especially completion or essay form. Another advantage is the complete objectivity of scoring, which is also true of multiple-choice form, but not of completion or essay. Still another is the suitability of it for many different kinds of material. Almost any facts fit easily into it, which is not so true of other objective types. Finally there is the doubtful advantage sometimes mentioned that true-false questions are easy to construct. If they are made by mere copying of appropriate statements from a book, with a slight change now and then, this may be correct. However, many writers have emphasized the difficulty of writing good true-false items.

Disadvantages often listed include the fact that true-false tests must be made much longer than other kinds of objective tests in order to reach a reasonable standard of reliability. Also they can seldom be made thoroughly diagnostic. They are likely to measure only simple recognition unless skillfully constructed. Also they encourage guessing unless a scoring correction for this factor is applied. The fact that they have appeared very infrequently in recent standardized tests of aptitude and achievement is evidence of the importance of these disadvantages in the thinking of leading psychologists and educators. However, their classroom use has continued in spite of these objections. One point against them which is seldom raised is the ease with which they lend themselves to cheating. Dishonest practices are absent altogether in some test situations, easily detected in perhaps a few, and cleverly disguised in some. Systems of signals for alternative response items are easy to invent, and will often distribute information widely without detection. Even ordinary copying, where seating is close, can go on quickly while the instructor and his assistant are not watching closely.

This chapter is not the place for an argument as to whether or not a teacher or civil service examiner should be a detective. Many opinions have been given regarding an examiner's responsibility in that regard, but here it is sufficient, perhaps, to point out that with careful preparation of alternate forms, cheating on true-false material can be reduced to a minimum in most cases.

Like any other form of objective test item, a series of true-false questions will need to be preceded by a clear set of instructions the first time it occurs in a test. How elaborate they need to be will de-

pend upon the previous experience of examinees for whom the test is intended. Illustrations or sample questions will usually be desirable unless the examiner knows that all individuals taking his test are accustomed to the particular system of recording answers that is required. Oversimplicity of directions with examples would be better than involved wording without illustration if there is any doubt about the ability of examinees to comprehend. The writer has found, among several formulations tried, the following most useful for lower educational levels:

Below are some numbered statements. Read each one carefully and decide whether it is true or false. If it is true, draw a circle around the T just to the left of it as shown below.

SAMPLE A: (T) F The capital of the United States is located at Washington, D.C.

If it is false, draw a circle around the F to the left of it.

SAMPLE B: T (F) The capital of the United States is at Boston.

Do not guess at answers you do not know. Guessing will count against your score.

For more sophisticated subjects, the following formulation was found by the writer to be quite satisfactory.

Draw a circle around T or F to indicate whether each of the following statements is true or false. Grading will be rights minus wrongs.

As mentioned before, the construction of true-false items is not so simple as it appears on the surface. The problem often overlooked is that of wording the statement in such a way that its content rather than its form will determine the response. If care is taken to solve this problem, the true-false item probably deserves a legitimate, though limited, place in objective testing, as Ross⁶ suggests. Ross points out its usefulness in "testing the persistence of popular misconceptions and superstitions." He also explains how it fits material for which an insufficient number of plausible alternatives can be found to make good multiple-choice questions. Four-choice items with two choices that are very readily eliminated from consideration because they are totally illogical might just as well have been true-false in form, since the chance of guessing the right answer

⁶ Ross, *op. cit.*

boils down to one in two. The number of plausible wrong answers available after a reasonable amount of careful thought would seem to be the logical determining factor in deciding whether multiple-choice form would fit or whether the true-false pattern would be more satisfactory.

Assuming that alternative response has been decided upon as the best in a given situation, the research worker will find some basic rules helpful. Each will be stated and discussed with examples.

1. *Statements must meet an absolute standard of truth or falsehood with only one interpretation possible; they must be entirely true or entirely false.* Ambiguous statements are unsatisfactory for this form. Also complex sentence structure often leads to one phrase being true while all the rest is false, which would be very confusing. A violation of the rule will be given with either a revision representing improvement or a totally new item that follows the rule.

VIOLATION: T F Human nature cannot be changed.

IMPROVED: T F Personality traits are seldom modified to a measurable degree by environmental factors.

The first example is a vague generality. Probably most writers would agree that basically little modification of human traits has been brought about through man's own efforts. Yet minor changes, and sometimes important ones, are accomplished by totalitarianism, psychoanalysis, counseling, religious fervor, etc. The issue is a controversial one with too many angles for a single question to cover it all. The second example is specific enough to be labeled false on experimental evidence.

Another illustration of this rule is needed.

VIOLATION: T F The Wechsler-Bellevue scale is the most widely used individual intelligence test for adults and is applicable to preschool children.

The part about its usefulness for preschool children is false, but the rest is true. It is two questions in one. Either part alone would be in accord with rule 1.

2. *The simplest, shortest statement that can be interpreted accurately is usually better than a more involved one.* Unless reading comprehension of complex sentences is the objective to be measured,

the language should be direct and simple, at the same time complete enough not to be ambiguous.

VIOLATION: T F Students who do not choose their own vocations, and whose parents project their own unrealized career ambitions upon them, quite unaware of the harm they may be doing, often do not have the basic aptitudes or interests that would be essential for success in these careers.

IMPROVED: T F Students whose parents have forced their own unrealized career ambitions upon them often lack the aptitudes and interests required for these careers.

To both of the above the answer is *true*. The first might very well have come verbatim from a textbook, but in any case it contains much unnecessary verbiage. One would not be safe in answering it without at least two or three readings, perhaps more. The improved version is simpler, more direct, and at the same time specific enough to be answered by anyone familiar with basic principles of vocational counseling. Logically this illustration leads to another rule.

3. *Modified wording is ordinarily better than verbatim quotation.* This point was discussed earlier and need not be elaborated further here except to say that a person who had memorized the example of violation of rule 2 from a book without grasping the central idea would probably be unable to answer the improved version. Unless rote memory is intentionally measured, the altered wording is more valid.

4. *Exact, quantitative terms, where possible, are generally superior to vague, qualitative ones.* This means, for example, that "300" would be better than "many" or that exact numbers of months or years would be more meaningful than such undefined phrases as "for a short time" or "over a considerable period of time."

VIOLATION: T F Successive administrations of the Stanford-Binet to the same individual may be expected to show very little variation in IQ up or down.

IMPROVED: T F Successive administrations of the Stanford-Binet to the same individual may be expected to show variations in IQ of 5 points up or down.

The answer to the improved version is clearly *true*. The interpretation of 5 points as "very little" variation might seem appropriate to the question writer, but someone else knowing that exact figure would consider it best described as "moderate" variation, and would perhaps answer the first version as *false*. Terms like "important" or "insignificant" may sometimes lead to the criticism, probably a just one, that success depends entirely upon how the question writer interprets the vague, undefined word. If the examinee can form a hypothesis, as he can after knowing a teacher for a while, his chances for passing are fairly good, but if the test constructor is unknown to those taking it, terms may have to be defined if there are two or more acceptable meanings.

5. *The meaning of a statement should be the only clue to the correct answer.* This rule has often been negatively stated thus: "Avoid specific determiners." Clues other than the ideas themselves may be derived from any of the following sources:

a. Strongly worded statements containing such words as *all*, *always*, *never*, *none*, or *nothing* have been found experimentally to be much more often false than true. Thus, if these words are frequently used in false statements and never in true ones, they become specific determiners. If their use is equally divided between true and false statements, they may be legitimate.

b. Such expressions as *sometimes*, *may be*, *often*, and *as a rule* are found much more frequently in true statements than in false ones. Careful balance between true and false statements containing any of these will avoid their becoming specific determiners.

c. Still another clue seems to arise out of an unconscious tendency of the test constructor to follow a rather consistent pattern in the answers. One such pattern discovered by the writer in a trainee in a civil service agency was F, T, T, F, T, T, F, T, T, etc. The trainee was totally unaware of any sequence to the answers until it was pointed out to him. Such a scheme will be discovered by an examinee who knows only a few of the answers. He will guess at the rest with fair to excellent accuracy, making use of the fixed pattern as a specific determiner.

d. Finally, any phrase frequently repeated and always or almost always associated with one answer throughout may serve as a clue. A reviewer may spot these more readily than the originator of the item can.

6. *Avoid double negatives.* One negative is clear, but two in one sentence are unnecessarily confusing. There is usually an easy way to correct these when found.

VIOLATION:	T	F	A counselor cannot adequately interpret the results of tests on which no norms have been established
IMPROVED:	T	F	Interpretation of test results can be adequate only when satisfactory norms have been established.

In reading the first example under the tension of a test situation, an individual can easily skip over the *no* without noticing it. This error might lead to its being marked *false*, while as it stands it is true. Such a careless mistake in reading is unlikely on the improved version, which is clearly true. In statements more involved than the examples above, the double negative may be more difficult to detect. Some test writers, when working out a single-negative statement that must be rather complex, will underline the word *not* if there seems to be a chance of missing the item because of a careless oversight. Such a practice probably will increase the validity of the item.

7. *False statements should represent an incorrect fundamental idea or principle, not a trivial detail.* The unfairness of so-called "tricky catch questions" has been frequently protested. Most such objections come from those who were looking for important ideas and not minor details of spelling, punctuation, etc. The usual answer to such arguments is that if any part of a statement is false, the entire statement is false, which may be correct enough, but still the question may not measure any significant teaching objective or job requirement.

VIOLATION:	T	F	The lower limit of dull normal intelligence is defined as an IQ of .80.
IMPROVED:	T	F	The lower limit of dull normal intelligence is defined as an IQ of 90.

In the first example, the inclusion of a decimal point before the IQ of 80 presumably makes the answer *false*, because an IQ of .80 does not exist. However, unless the item is intended to measure ac-

curacy of proofreading, which is unlikely, the examinee has a mental set or readiness for ideas, not mechanics of printing. The improved version would detect a confusion in thinking about categories of intelligence and would not unduly penalize the rapid, somewhat careless reader. Minor changes in spelling of well-known proper names and the substitution of B.C. for A.D. after a historical date would probably be considered an illustration of the same fault.

8. *Authority must be mentioned for statements that might be considered controversial.* If statements of hypotheses or theories are used for true-false items, the reader must be informed specifically what theory or whose opinion the writer had in mind. In academic courses following one textbook closely, it is often assumed that answers are to be on the basis of what the textbook says, not from any other readings or from the examinee's own thinking. Under these conditions, the test constructor obviously does not need to waste space beginning every true-false statement with, "According to _____ (author)," since this viewpoint is understood from the beginning. However, there are many teaching situations in which no such assumption can be made, and the same applies to practically all personnel-selection written tests. Usually a matter of opinion or theory is good in true-false form only if it is widely accepted or if its author is well known.

VIOLATION: T F The limits of borderline intelligence are 70-80 IQ.

IMPROVED: T F According to Terman the limits of borderline intelligence are 70-80 IQ.

or

T F On the Stanford-Binet scale the limits of borderline intelligence are 70-80 IQ.

To a writer whose knowledge of testing is narrow, the first of the above examples would seem satisfactory. To a student who had confined his reading on the subject to one or possibly two books, there could be no possible correct answer but *true*. To the student who had encountered the fact that Wechsler defined borderline intelligence as extending from 65 to 80, the first example would be puzzling. According to Terman, the limits given in the question would be correct, while according to Wechsler they would be false.

More information would be needed to make it a fair question. The second and third examples are specific, and the additions do not give away the answer.

Statements of fact well established by generally accepted proof will need no authority, but in many fields of study, theories and even hypotheses are important. In most instances, these are clearly distinguishable from fact by any person well acquainted with the subject matter, but a person less broadly informed may make a serious error on this point. Philosophy, psychology, sociology, economics, education, and even the "exact" sciences of physics and chemistry are full of controversy, as are many other subject-matter areas. Easily overlooked may be items like the one below:

VIOLATION:	T	F	The mind is divided into two separate and often conflicting parts: the conscious and the unconscious.
IMPROVED:	T	F	Psychoanalytic theory divides the mind into two separate and often conflicting parts: the conscious and the unconscious.

The improved version is true, but there are many who would argue with the first example on the basis that the parts are not separate and distinct and that the arbitrary division is a distortion of observed facts of behavior. Therefore the first version is unfair as a test question.

With these eight rules, one should be able to make true-false questions deserve a place in objective testing which some critics would deny them. Not every fault can be avoided by applying these simple principles, but many poor items now in use could by means of them be improved. Much of the research which has turned out unfavorably in regard to true-false form may have been done with faulty items. The writers usually do not give the reader any detailed evaluation of the material used. ✓

Multiple-choice Items. The factor of chance success is reduced when, instead of two possible answers, there are three, four, or five. If there are no absurd, too obviously wrong answers included, the chances of guessing the correct answer to a totally unfamiliar question would be, in the case of three possibilities, one in three; four choices, one in four; and five choices, one in five. This theoretical

assumption concerning chance success cannot safely be made, however, unless wrong choices are about equally plausible, which is probably seldom the case. The likelihood of correct guesses is usually only somewhat less than it would be in alternative-response or two-choice forms.

A multiple-choice item is an incomplete statement with three, four, or five possible completions, one of which is either correct or at least the best among those given. Or it may take the form of a question with one correct or best answer among three to five given. The incomplete statement or question is known as the premise, and the answers from which one is to be selected are called choices. Wrong answers are sometimes referred to as distractors. The usual number of choices is four, and they may be numbered or lettered (1, 2, 3, 4, or A, B, C, D). Beginning each choice on a new line seems better, from the standpoint of ease of reading, than running all the choices in together. Quick location of each choice for comparison with others is desirable.

Punctuation has varied from one agency to another, but the simplest appears to be best. Use of the semicolon after all choices except the last, which is followed by a period, is widely accepted practice, but seems unnecessary. A colon after the premise, provided that a question mark is not obviously required, seems not to be compulsory according to any rule of punctuation known to the writer. Only one phrase is needed to complete the premise, and without the insertion of letters and wrong choices, the sentence would be correct without such punctuation. Similarly, the placing of parentheses around the numbers or letters of choices appears to be nothing more than two unnecessary strokes on the typewriter. Economy of labor for the writer and reader would seem to suggest the acceptance of one or the other of the forms below or some combination of the two.

SAMPLE:

1. A test that predicts how quickly an individual could learn a skilled trade would measure
 1. achievement
 2. general intelligence
 3. mechanical aptitude
 4. arithmetical reasoning.

2. If a test predicts how quickly an individual could learn a skilled trade, what would it measure?
- A. General intelligence
 - B. Achievement
 - C. Mechanical aptitude
 - D. Arithmetical reasoning.

An additional point in regard to form involves the way in which answers are to be recorded. Underlining takes more time to grade than writing of a number or letter on a line before or after the question or on a separate answer sheet. Machine-scored answer sheets have special instructions, mentioned earlier, but hand-scored tests might bear instructions like the illustration below, which proved to be the best of several tried by the writer in academic and public personnel settings:

INSTRUCTIONS: The following statements are incomplete. After each statement you will find four or five words or phrases that might be used to complete it. These words or phrases are lettered A, B, C, D, and sometimes E. Find the word or phrase that will make the statement true, or most nearly true, and put *its letter* after the proper number on the answer sheet. Look at the sample below:

SAMPLE: The capitol of the United States is located at

- A. New York City
- B. Washington, D.C.
- C. San Francisco, California
- D. Chicago, Illinois.

Washington, D.C., is the correct answer, and B has been placed after SAMPLE on your answer sheet. Remember that your answer to each question is *always only one letter*.

Persons accustomed to multiple-choice tests do not need such elaborate detail, but for a first experience with this form it is not needlessly long. If simplified instructions will be entirely clear to all examinees, since they have all answered multiple-choice items before, the writer of the examination might use such a simple sentence as, "Write the *letter* of the best answer on the line to the left."

Multiple-choice items need not meet the absolute standard of truth and falsehood demanded of true-false material. Often the multiple-choice question may call for "the best of the following

methods," etc., where one could not be sure that any of the choices given is a perfect solution. The phrase "of the following" is a convenient way out of argument from perhaps a really brilliant, exceptionally well-informed individual who may think, and perhaps justifiably, that he has a better solution to the problem than any of the four stated in the choices. Not every question will call for such a phrase, of course.

One of the principal advantages of multiple-choice items is the wide variety of materials which can be readily used in test form. Many tests are too restricted in the structure of this kind of question, and follow rigidly some fixed pattern instead of making use of the great flexibility that is possible. A number of studies cover various aspects of the possibilities of this form. Mosier, Meyers, and Price⁷ propose 14 types of questions that can be fitted into this pattern. The list is not entirely complete, and it does not intend to prescribe the exact language to be used, but it is proposed as a guide in formulating a wider variety of questions than will usually be found on any one test.

A somewhat edited version of this list will be given here with illustrations. Originality will be required to convert these examples into other areas of knowledge, but careful study of this outline and the illustrations should broaden the scope of the research worker in test development, enabling him to develop greater flexibility in the use of what is perhaps the most valuable tool in objective testing—the multiple-choice form. The questions stated under each heading are among the best tried by the writer on elementary psychology students. They have all been evaluated by careful item-analysis procedures to determine whether they are appropriate in difficulty and whether they show at least fair discrimination between good and poor students in the course as a whole. The sizes of the groups varied considerably for the different items, but all were good.

General categories are underlined. Answers are shown at the right.

1. Definition:

An eye defect characterized by irregularities in curvature of the cornea or lens is

⁷ Mosier, Charles I., Meyers, M. Claire, and Price, Helen G., "Suggestions for the Construction of Multiple-choice Test Items." *Educ. Psychol. Measmt.*, 1945, 5 (No. 3):261-271.

1. myopia
2. hyperopia
3. astigmatism
4. strabismus.

3

Astigmatism is best defined as which of the following?

1. The eyeball too short, causing images to focus behind the retina
2. The eyeball too long, causing images to focus in front of the retina
3. An irregularity in curvature of the lens or cornea
4. A blind spot.

3

2. *Purpose:*

A control group is used in order to

1. hold all factors constant as far as possible except the variable being studied
2. increase the population studied and make the experiment appear more impressive
3. compare two entirely different samples of the population
4. none of the above.

1

3. *Cause:*

The causes of feeble-mindedness include

1. brain injury
2. heredity
3. diseases of the brain
4. all of these.

4

4. *Effect:*

Deficiency in thyroid secretion from birth will produce

1. mongolism
2. cretinism
3. hydrocephaly
4. microcephaly.

2

5. *Association* (similar to cause and effect, but no direct relationship):

Auditory hallucinations usually occur with

1. manic-depressive psychosis
2. schizophrenia
3. neuroses
4. chronic alcoholism.

2

6. *Recognition of error:*

Which of the following statements is *false* when applied to preschool children?

1. Complex group cooperation is seldom observed in spontaneous play behavior.
2. Giving affection freely is essential to normal personality growth.
3. Imaginary playmates are a normal phenomenon often observed.
4. Fantasy concerning wild animals related as actually seen indicates the beginnings of a pathological trend.

4

7. *Identification of error:*

Several rats learned to turn to the right and go toward a green light and to avoid a red light that appeared on the left. The experimenter concluded that the animals could discriminate between red and green.

This conclusion

1. is entirely justified from the results, since the rats all learned to respond correctly
2. is not necessarily correct, since not enough rats were used
3. would not necessarily be correct, since the animals might have learned a position habit
4. might be subject to question if on further trials an occasional error were made.

3

8. *Evaluation:*

The recitation method of learning is

1. superior to many rereadings for most kinds of material
2. wasteful, because errors in reciting are firmly fixed
3. about equal to rereading in bringing results promptly
4. not recommended by most educational psychologists.

1

9. *Difference:*

Of the following the *chief* difference between amentia and dementia is that

1. the former refers to children, while the latter applies to adults
2. amentia is lack of normal mental growth, while dementia is loss of mental functioning once acquired
3. the former is greater in degree than the latter in most patients
4. amentia requires hospitalization, while dementia does not.

2

10. *Similarity:*

The basilar membrane and hair cells in the inner ear and the retina of the eye are alike in that both

1. have the same cell structure
2. convert stimuli into nervous impulses
3. respond to light waves
4. interpret objects in the environment.

2

11. *Arrangement:*

If the following methods of measuring retention were arranged in order from that indicating the greatest amount retained to that showing the least amount remembered, the correct order would be

1. recall, relearning, recognition
2. recall, recognition, relearning
3. relearning, recognition, recall
4. relearning, recall, recognition.

2

12. *Incomplete arrangement:*

In the following series, which category has been omitted: idiot, imbecile, moron, dull, average, superior, very superior, near genius or genius?

1. Borderline
2. Feeble-minded
3. Mental defective
4. Ament.

1

13. *Common principle:*

Which of the following does *not* belong with the others?

1. Kohler
2. Koffka
3. Wertheimer
4. Adler.

4

14. *Controversial subject:*

Those who advocate the use of projective techniques generally agree that

1. the TAT is much more valuable than the Rorschach
2. questionnaire-type personality tests would be better if those taking them would answer truthfully
3. many projective methods have already reached an advanced stage in standardization

4. the Rorschach is of value for personality structure, while TAT reveals content. 4

15. *Multiple answer:*

Slow speed of reading may be caused by

1. too many fixations or pauses per line of print
2. pauses that are too long in duration
3. regressive movements of the eyes to pick up what was not perceived at the first fixation
4. all of the above. 4

Comments on these categories and examples may help to make them clear. The reader will do well to bear in mind that in each case there are several variations from the wording given here that are acceptable as belonging in the category.

1. *Definition:* As shown under this heading, the term may appear in the premise with its correct definition among the choices. However, many acceptable items have the definition in the premise to be matched with the correct term among the choices. Still others have an entire sentence in technical terminology in the premise to be matched with the same law or principle stated in more everyday language in the correct choice. The distractors in this case may state some popular misconception about the law or rule stated in the premise. Such exercises in recognition of the central idea or thought expressed in different words can often screen out the rote memorizers. To them there is nothing in common between two statements of the same problem unless they contain essentially the same words.

2. *Purpose:* This group takes care of many questions that begin, "Why is this done." It may also include the selection of the most important reason from among some trivial reasons.

3. *Cause:* The example given fits equally well under 15, since it requires recognition of several causes, but for other material only one cause among those stated could be correct. Other items in this group may begin with a premise such as, "Under which of the following conditions would this happen?" The choices would then state various sets of conditions, only one of which could be right.

4. *Effect:* This is the reverse of the above. The cause is now stated in the premise, and the result, or most likely result, must be picked from the choices.

5. *Association*: The kind of example used here, connecting symptoms with diseases, is quite common. Here the cause-and-effect relationship is not very clear, but the two things seem to occur together. Sometimes the connection is merely one of space or time.

6. *Recognition of error*: The example is almost a modified true-false pattern. In this group may be found such wordings as, "Which of the following is *not* used?" This group is adaptable to mathematical equations, administrative procedures, sentence structure, etc.

7. *Identification of error*: This group goes one step further than the above. It calls not just for evaluation of something as incorrect, but for selection of the principle which is violated. The error must be given a name, a descriptive term from among several choices.

8. *Evaluation*: In addition to the form given in the example, the question may call for the best of four or five methods of achieving a specified purpose or goal. Ordinarily the mental process called for is reasoning, not just memory of the opinion of any particular author. Accurate memory of the content of reading matter on the subject may be helpful in furnishing a basis for thinking, but clearly organized thought will usually be a minimum requirement for correctly answering items in this group. The same may be said for many items in each of the other groups, although people will differ in their approach to these. Some will be inclined to depend entirely upon memory for what they have read or what they have been told, while others habitually solve as many problems as they can by common sense. Either method may be successful in many cases. The well-trained, natively versatile thinker will try both approaches. If one will not work, the other usually will.

9. *Difference*: Underlining the word *chief* is important when the distractors are real but unessential differences. If only one of the choices is acceptable as a difference, the word *chief* may be omitted.

10. *Similarity*: Here again organized thinking is likely to be more useful than mere memory for reading matter or lecture.

11. *Arrangement*: This illustration alone will answer quite well the argument advanced by some teachers that objective tests do not measure the student's ability to organize his thinking. The essay test may go somewhat beyond this group of objective items in requiring clear expression, but placing steps in their order according to a cer-

tain rule or principle is a minimum essential for answering material in this form. Application of a rule or law to a practical problem may be demanded by an item in this group. In such a case, rote memory for verbatim quotations is not apt to be of much help.

12. *Incomplete arrangement*: This is a somewhat less useful and more complicated variation of the above. The illustration does not inform the reader where the insertion needs to be made, and if it did so, the item would probably be easier (or too easy).

13. *Common principle*: The example given is a rather widely used pattern, but the category includes also a premise listing several ideas or objects and choices including one principle common to all the list.

14. *Controversial subject*: The illustration eliminates possible argument on the part of the examinee who does not agree with the school of thought whose opinion is stated. Many educators and psychologists do not believe that projective methods are sufficiently objective to be of value, but it is probably important for students in these fields to know what has been said in their support. The arguments for and against a contested theory are significant in many subjects, and they can both be required on a multiple-choice examination.

15. *Multiple answer*: This group overlaps several of those previously mentioned. The illustration is a special case in the *cause* category, requiring somewhat more careful study than an item calling for a choice of only one idea instead of more than one that could be correct. There has been some argument among those doing research in objective-test development as to whether an "all of the above" or "none of the above" choice should ever be included as a final one. Either of these phrases could become a specific determiner unless it were as often an incorrect answer as a correct one. An item could be designed so that the first choice was the only right one and the final "all of the above" would be wrong. Similarly, "none of the above" could be incorrect. The writer is of the opinion that, in the example given for this category, the selection of choice 1 without recognition of 2 and 3 as true would indicate incomplete learning or careless thinking, or perhaps both. There is nothing inherently unfair about such questions once the student or job applicant becomes accustomed to the mental process involved in solving this

kind of problem. A single sample included in the instructions with clear explanation would suffice.

F. L. Ruch made abundant use of this last group in his workbook mentioned earlier in this chapter. His third and fourth choices are often "both of these" and "neither of these," referring to choices 1 and 2. In such instances, his choices 3 and 4 are as often wrong as right, approximately, and cannot therefore be considered specific determiners. Most of the objections to his items of this type observed by the writer have come from inferior students. The use of such phrases as "both of the above" and "neither of the above" is a convenient method of handling the situation frequently encountered by the test constructor when he simply cannot find or originate more than one or two plausible distractors. Overuse of this device would be rather undesirable as a substitute for good distractors.

In presenting rules for writing multiple-choice items, an attempt has been made to ensure as far as possible their validity. Also the basic principles that will provide for specific and uniform interpretation have been included. Wherever possible, positive rather than negative formulation of the rule has been attempted.

1. *The correct choice should usually be about the same in length as the distractors, not consistently longer or shorter.* If this precaution is not taken, the length of the right answer becomes a specific determiner.

VIOLATION: The present trend in vocational counseling is toward

1. increased use of interviews and other exploratory techniques besides tests
2. administration of a wider variety of tests to each person
3. a study of aptitudes rather than interests
4. more emphasis on general intelligence. 1

This would in itself be an excellent item, but if in all or most of the surrounding material the right choices were the longest, there would be a decided violation of the above rule. If this example had to be altered to correct the possible specific determiner, the correct answer could read like this:

IMPROVED: 1. increased use of other techniques besides tests.

If the correct choice would not be clear when shortened, as is true in some instances, the defect could still be remedied by adding a few appropriate words to one or more of the distractors.

2. *The correct choice must be different from the distractors consistently in meaning only, with no superficial verbal clues.* Consistent grammatical differences between the correct and the wrong choices may creep in without the test writer being aware of this clue. It may amount to the repeated selection of the infinitive form of the verb on the right choice and the participle form on the others. Sometimes the mere word usage gives the answer away.

VIOLATION: Psychiatrists usually evaluate personality by means of

1. inventory tests like the Bernreuter
2. the Rorschach
3. the Thematic Apperception Test
4. the psychiatric interview.

4

In this example the inclusion of the word *psychiatric* gives the answer away. Other instances of this clue are more subtle, but even obvious violations are not uncommon.

3. *If an item depends in any way upon the preceding one, neither must reveal the answer to the other.* Items will ordinarily be independent of one another, but sometimes a central problem can be followed effectively through two or more stages with more than one item. When this is skillfully done, there are no clues, but the common error illustrated below is to be avoided.

VIOLATION: The part of the brain chiefly concerned in higher mental processes is

1. the cerebellum
2. the cerebrum
3. the medulla
4. the midbrain.

2

According to the mass-action theory, the cerebrum

1. is highly localized in function with regard to higher mental processes
2. acts as a whole in higher mental functions
3. is not localized in sensory functions
4. is not localized in motor functions.

2

These two questions seem satisfactory if they are not immediately adjacent to each other. The remedy might be to insert five or six other questions between them. Legitimate dependence in a sequence of items is shown by the following:

IMPROVED: John, a college freshman, is a very slow reader, even though he has very superior intelligence. Of the following, the most likely cause is

1. faulty perceptual habits
2. poor outlining of chapters so that important facts are missed
3. cloudy thinking and poor comprehension
4. a poor memory for facts. 1

A good diagnostic test for the above case would be

1. Wechsler-Bellevue
2. Minnesota Test for Clerical Workers
3. the SRA Reading Record
4. a reading-readiness test. 3

The case description in the premise of the first item is essential for answering the second, but there is nothing in either that would make the answer to the other obvious. If the first sentence in the premise to the first one were added at the beginning of the second one, either could be used alone with the level of difficulty remaining the same for each. In the first pair, however, the word *cerebrum* in the premise of the second item suggests the answer to the first. This kind of a connection between successive items is often more subtle than in the illustration here, but still effective in making one of them easier.

4. *Answers should follow a random pattern.* They may give a clue if too regular. This problem has been discussed under true-false material, but here something new may occur. Unconsciously many test writers give the correct answer to choice 2 and 3 more often than to 1 and 4. Also some unique patterns have appeared on the scoring keys of some tests observed by the present writer, such as A, D, C, B, A, D, C, B. A person with a strong "gambling instinct," or, more scientifically, above normal willingness to take risks, might hazard a guess on many items of which he knew nothing and succeed on the basis of a pattern followed fairly consistently throughout.

5. *There should be a clear central problem in each question.* A multiple-choice item should not be a mere series of unrelated true-false items, but it should have a premise that sets forth a definite problem, the solution to which is to be found among the three to five choices.

VIOLATION: Vocational guidance

1. is of no value for college freshmen
2. always tells a person the best occupation for him
3. goes much further than the mere administration, scoring, and interpretation of a battery of tests
4. consists of the evaluation of basic aptitudes of the individual in an effort to redirect his interests along new lines.

3

IMPROVED: The present trend in vocational counseling is toward

1. increased use of other techniques besides tests
2. administration of a wider variety of tests to each person
3. a study of aptitudes rather than interests
4. more emphasis on general intelligence.

1

The first example is inferior to the second in that the premise does not set forth any definite problem, but only a general topic, on which one true and three false statements follow. The improved version narrows the subject down to a specific point, and each choice must be related to that one point, the present trend, in order to arrive at the correct answer.

This rule needs further illustration, since sometimes a very vague premise makes the interpretation of the problem anything but uniform.

VIOLATION: A student can learn best by

1. short periods of distributed practice
2. rather long periods of study or practice
3. rereading several times
4. observing rather than doing.

IMPROVED: A student can best learn complex skills by, etc.

The original premise was incomplete and vague, since it did not state what kind of material was to be learned. Unless the scope of the problem were limited, no correct answer could be given, since

for skills choice 1 would probably be true, while for some kinds of meaningful material in large units, choice 2 would be more correct. When the premise becomes precise, as in the improved version, an answer can be justified.

6. *Wording should be as concise as possible without failing to explain each important point clearly.* There are several implications in this formulation. First, the research worker may infer that no irrelevant information belongs in the premise or choices, since it would add nothing to what is measured by the item and would unnecessarily lengthen and complicate the reading.

VIOLATION: John, nine years old, in the fourth grade in a small rural school, made an IQ of 95 on the Stanford-Binet. This would classify him as

1. dull
2. genius
3. defective
4. average.

4

Information about the kind of school he attended might be of value in a complete study of his case, but it is of no help in answering the above question. Neither is his grade placement. The question is otherwise practical and sound. Improvement would result from eliminating the useless facts from the premise.

A second inference would be that the use of too much verbiage to express one idea may easily change what is measured by the item from information or aptitude to reading comprehension at a high level of complexity. In the writer's personnel-selection experience, it has been his frequent observation that some professional social workers and educators are particularly guilty of this offense. Their items frequently contain good ideas, and the test technician would often prefer to have them write the questions in preliminary form. When boiled down to essentials, they often are recognized by the original writers themselves as improved, and are accepted by the trained psychologist as entirely satisfactory material from the construction standpoint. An illustration is given below:

VIOLATION: The general aims to be accomplished in the elementary school as viewed by the teacher herself are very important to the progress of education in that they should stress primarily

1. that the adjustment and welfare of the group should take precedence over the problems of the individual child
2. that the group activity should in most instances be subordinated to the needs and problems of the individual child
3. that proper balance be maintained between group needs and individual needs in order that the maladjusted youngster not be forgotten, and yet that all shall be trained in the cooperation necessary for social enterprise
4. the curriculum, in order that the subject matter will be presented in the most pedagogical and logical manner that is understood by teachers trained in the objectives of modern progressive education.

3

Every profession has its jargon, and the test technician would be making a serious mistake who tried to discourage the specialist from using it, but when mere verbosity seems to have been set forth as a virtue, as in the above instance, the validity of the item may be questioned. The educator who originated the above material had in mind a very good problem, but several readings would doubtless be required to grasp it. The meat of the matter could perhaps be set forth in language entirely acceptable to the profession and quite clear to the informed reader in something like the following manner:

IMPROVED: In the elementary school classroom situation, the teacher should stress primarily

1. group needs rather than the needs of the individual child
2. needs of the individual child rather than group needs
3. the proper balance between group needs and individual needs
4. the subject matter and its presentation rather than satisfaction of the needs of children either individually or collectively.

3

There are perhaps a few exceptions to rule 6 which should be allowed. The ability to pick out the facts that apply to a given problem from a mass of data, much of which does not apply, may be said to be an indication of aptitude for specific types of work involving problem solving. Practical situations frequently confront us with irrelevant information, and the intelligent individual will cast this aside, not attempting to use it. Problems in arithmetic reasoning may be found in standardized tests and in books which call for this

selectivity. In several scientific fields, such problems have been useful. The following might clarify this exception:

ACCEPTABLE: Four clerks were hired at 80 cents per hour to check and file 7,680 forms. They worked at an average speed of 40 forms per hour, and completed the entire project in 6 working days. At the same working speed, how many clerks would be required to complete the same amount of work in 3 working days?

1. 2 clerks
2. 8 clerks
3. 12 clerks
4. 7 clerks.

2

The nonessential facts in this problem include (a) the pay, (b) the number of forms to be completed (a filing project would have been sufficient), and (c) the working speed per hour (as long as a fairly uniform speed is assumed). Omission of these data, however, would have made the problem easier and would have carried it over into the academic rather than the practical. The wrong choices have been worked out on the basis of frequent errors made on the part of careless thinkers who go into the problem in a trial-and-error fashion. Choice 1 would be selected as correct by anyone who happened to notice that 3 days is half of 6 days, and concluded that half of 4 would be 2. Some would try to multiply two figures that were handy, like 4 clerks times 3 days, giving 12 (choice 3). Thus the data not needed for the solution serve as confusing factors, which are not unfair from the standpoint of likeness to real situations.

7. *Information given in the premise must be complete enough to make one answer justifiable.* Through an error on the part of the item writer, credit sometimes has to be given for two answers neither of which can be proved wrong. This situation should be foreseen and avoided as much as possible. The cause of two choices turning out to be right is often found in insufficient information in the premise to establish beyond doubt the intended right answer. To achieve the goals set by rules 6 and 7 will require practice. The happy medium between the extremes of excessive detail and excessive brevity is not easy to attain in wording some kinds of problems.

An item with two or more possible answers because of insufficient information in the premise is given below.

VIOLATION: A child's IQ improved from 80 to 98 in two years. The examiner could best conclude that

1. the first test was given by an inexperienced person
2. the child's environment had improved
3. the second test was too high
4. there was an error in scoring in one test or the other.

There would probably be differences of opinion as to whether choice 1 or choice 2 would be the best answer, and unless further information were available, neither point of view could be established with any certainty. If the point could be briefly stated in choice 1 that inexperienced examiners often do not get the best cooperation from the child, this choice would be improved, but it is possible that another choice requiring a briefer statement would be better. The premise needs to set forth some indication of the environmental forces that may have been operating during the two-year period. Without this addition, the possibilities would be so numerous that the problem would be unsolvable. Therefore the following is suggested as one of several acceptable revisions that might be made:

IMPROVED: A child was removed from a very poor institution to an excellent foster home, with a resulting improvement in IQ over a two-year period from 80 to 98. Of the following, the *most likely* conclusion is that

1. on the first test unfavorable environmental factors prevented functioning up to maximum capacity
2. the first examiner was careless about scoring his test, and made some trivial error in arithmetic
3. the second test was too high
4. no test is accurate on a child away from his own home.

1

There is now a defensible answer. Choices 2 and 3 are possible, but not likely. Probably the question could have been satisfactorily revised without doing as much to the choices as was done here. Addition to the premise was the most necessary change.

How exact or elaborate an item needs to be will depend in part upon how relatively naïve or sophisticated those taking the test are. Multiple-choice items are often reported to be more difficult for the student who is widely read in the field and who depends upon his own thinking to a large extent than for the one who has information limited largely to one textbook and depends largely upon rote memory for that. The former can question the correctness or finality of almost any given choice unless it is qualified by numerous phrases stating exceptions, etc. The latter quickly and uncritically selects what he has read or what he has been told. Hence the mediocre student may answer "correctly" in an uncritical manner an item which is missed by the outstanding student, who would like to argue the matter with the question writer, probably with much justification.

For beginners in any subject, or for routine jobs, the multiple-choice items may be quite naïve and brief. Probably they should be, in order not to create unnecessary confusion. For graduate students, or for highly technical or administrative personnel, valid questions will have to be much more exact and complete. Thus rules 6 and 7 will have to be applied in a flexible fashion, with the situation and purpose of the test largely governing the test writer's formulation.

8. *Distractors must be plausible.* In an open competitive examination for Park Keeper in a large city appeared the most absurd disregard of this rule ever encountered by the writer.

VIOLATION: While on the job a Park Keeper should not

1. cuss
2. chew
3. smoke
4. go to sleep.

4

The entire item is foolish and too obvious to be of any value, although some would question whether number 1 should not be regarded as just as good as 4 for an answer. Sometimes the distractors are not foolish, but merely unrelated to the premise. The following example is one that can be answered correctly by someone who has only a very vague idea of what sublimation really is.

VIOLATION: An example of sublimation is

1. binocular disparity
2. retroactive inhibition
3. creative art
4. multiple personality.

3

The distractors are so implausible that they can be eliminated easily by a person who grasps their meaning only vaguely. After all, it is not the meaning of these wrong choices that the question is intended to require, but the concept of sublimation, the right answer, that must be clearly understood. To ensure the validity of the item, that it measures what it is intended to measure, the distractors must be more closely related to the problem. They must not be merely other concepts in the general field of psychology, but they must be close to, yet not exactly, correct. They must seem right to the person who does not really know the subject matter, even though he may be rather clever. No common-sense judgment can be allowed to replace the knowledge that the question calls for if the item is to be valid. One possible revision is:

IMPROVED: An example of sublimation is

1. creative work in some branch of art
2. telling smutty stories instead of expressing sex directly
3. temper tantrums
4. picking a fight every time there is a difference of opinion.

1

All of the wrong answers here refer to forms of behavior that express something in the personality of the individual. There does not seem to be any chance of using some method other than the correct information here. There are no clues, and a negative approach by eliminating wrong choices does not appear to have possibilities.

Rule 8 is not one of the easiest to follow. Plausible alternatives take resourcefulness of imagination to find or to originate. Yet if it is not followed, guessing right answers becomes increasingly easy. No further instructions can be given for finding the reasonable choices except that facility in thinking in the multiple-choice pattern comes with practice to anyone with at least a moderate degree of word

fluency. If a person lacks even an average amount of this native aptitude, he will probably never develop into a very good test constructor.

9. *Vocabulary should be appropriate to the group for which the test is intended.* For academic courses the wording can be technical, and on some jobs the competitive tests can be. For jobs requiring little formal education, the tests for personnel selection need to be in vocabulary that can be read and understood by persons whose grade achievement equals the average of acceptable applicants.

In this connection the writer recalls an examination given by a public personnel agency for a position as chief cook in an institution. A Negro was wanted who could supervise several subordinate cooks. The general style of the test material is illustrated by the following:

VIOLATION: What methods do you use in orienting new employees under your supervision?

1. Give them an over-all picture of the duties and responsibilities.
2. Let them work out their own problems from the very start.
3. Assign one simple task until it is learned, then go to the next, etc.
4. None of the foregoing is correct.

1

The above deals with a generally accepted practice in supervising new employees, and it might apply well enough to the job in question. However, the average applicant had little verbal capacity and needed very little to perform the duties. Furthermore, acceptable work of this type could be and often is done by persons of fifth-grade achievement or even lower. The vocabulary of the item might be appropriate for high school graduates, but it could be nothing more than confusing to most employees in positions of this sort. Use of rather simple and even colloquial phrases would meet with no criticism in this setting from anyone interested primarily in the validity of the test. The following revision was suggested by a reviewer.

IMPROVED: How do you break in a new cook?

1. Show him where things are and tell him what are the most important things you want him to do.
2. Let him find things for himself and go ahead on his own.
3. Tell him only one thing at a time until he learns.

4. Make him understand that he must obey your orders at all times or you will see that he is fired from his job. 1

This wording is straight to the point for anyone with a minimum of formal education. If the slang can be excused, the only criticism of this revision is probably that in the case of some mentally defective subordinates, the supervisor will be obliged to do choice 3 rather than choice 1.

The choice of vocabulary appropriate to a low educational level may be aided by the use of various vocabulary lists published to show the frequency of use of different words. Usually, however, common sense will be all that is required to solve this problem. However, the choice of language appropriate to a job may be a more difficult problem than the example above would indicate. Review of items by someone experienced on the particular job in question may result in the change from academic terms to the everyday usage of the employee that will result in the material making sense to the applicant. Without such a change the item may lack validity. A well-qualified worker may miss the point altogether.

10. *Facts called for should be useful ones.* Memory for minor details that will never be applied, or that an individual can look up easily on those rare occasions when needed, would involve a needless expenditure of time and effort. If a teacher requires his students to "clutter up their minds" with trivialities, he usually becomes quite unpopular. The arithmetic teacher who required her pupils to memorize the fact that there are 63,360 inches in a mile did not think through very far on her teaching objectives, since one would seldom if ever have to reduce inches to miles or miles to inches. If such a problem were encountered once or twice in a lifetime, it would be better to reduce to feet, then to inches, than to burden one's mind with such figures. The test technician will find it unnecessary to memorize every statistical formula. He will need to know, however, where to find the formula for any problem he proposes to do. He must then be able to use it and to interpret what he gets. The following is a simple illustration of useless facts that a student working with intelligence tests could find in his book when he is scoring.

VIOLATION: The number of months MA credit for each item passed at the Superior Adult I level on the Revised Stanford-Binet is

1. 5
2. 4
3. 6
4. 10.

 2

The remedy is obviously to eliminate this question altogether and find some point worth learning about which a question would be justifiable. There are some exact figures worth remembering in the field of tests and measurements, such as the limits of borderline intelligence in IQ. This could be looked up, but it is used often enough by those giving and interpreting tests to make accurate memory for the limits an actual timesaver. In statistics, the student who remembers that approximately 68 per cent of the cases in a normal distribution fall between 1 sigma above the mean and 1 sigma below the mean probably has had considerable experience in working problems in this field. Only subject-matter experts can judge wisely what facts are worth learning and what are not. Here is a useful problem in selecting the best test to apply.

IMPROVED: The intelligence test considered most appropriate for individual administration to adults is

1. the Stanford-Binet
2. the Wechsler-Bellevue
3. Otis Self-Administering
4. the Henmon-Nelson.

 2

Review by the subject-matter specialist may usually be the best criterion of usefulness of facts incorporated into a multiple-choice test.

11. *Avoid confusing sentence structure.* There seems to be no way of stating this rule except in the negative, since it covers literally a multitude of sins. One of these is stating some of the choices in the negative and some not when the premise is essentially a negative. This is a subtle form of double negative, much less obvious than the one considered earlier. Study carefully the following:

VIOLATION: Of the following steps in test standardization, the one most often neglected is

1. finding the mean of the group
2. not furnishing adequate evidence of validity

3. finding the standard deviation of the distribution of scores obtained
4. not trying out each item carefully before compiling the final form.

 2

Omission of *not* at the beginning of choices 2 and 4 would eliminate a source of considerable confusion. Another kind of confusing structure might be illustrated in the same item if choice 1 began with "to find" and choice 3 with "finding." This is less serious, but on the whole it is better to be consistent in using either the participle (finding) or the infinitive (to find) for all choices.

The final test of the merit of a multiple-choice item is always whether it makes good sense to a person who should be prepared to answer it. If it is vague, ambiguous, tricky, unnecessarily involved in structure, or trivial, it will not measure up to an acceptable standard. If it is too obvious or too much a matter of one relatively unknown man's opinion, it is useless. There may be other forms of undesirable material not mentioned in this section, but these rules seem from wide experience to be the ones most often violated. Careful review and revision will usually result in worth-while correction.

Completion Test Items. Of somewhat more limited usefulness than the multiple-choice form is the completion item. This kind of question consists of a sentence or paragraph in which one or more words or short phrases have been omitted. Blanks are supplied either in the body of the material or in the margin to be filled in by the examinee. The instructions are usually very simple, like, "Fill in the missing word or phrase in each blank." Completion items are particularly valuable in mathematics or in the physical sciences where a problem is presented requiring computation or in any field where isolated bits of information need to be recalled.

Among advantages may be mentioned that recall is demanded instead of mere recognition. Study habits must be thorough, therefore, in order for recall to take place. Guessing, which is likely on true-false and possible on multiple-choice material, is reduced to the very minimum. On the other hand, there are limitations that deserve mention. Unless they are very carefully constructed, completion items are likely to require no mental processes higher than rote memory. They measure detailed factual information largely, not

thinking or organizing ability. There are some rather subtle difficulties in construction, which, if not observed and resolved, will lead to lack of objectivity in scoring. Grading is not nearly so rapid as it is for other types of objective questions, and may often call for expert judgment rather than clerical help, since acceptable answers not included by the writer in the scoring key are apt to occur. Objectivity may often be much less than in other types of so-called "objective" test items. In this respect, completion material falls somewhere between the completely objective and the essay forms. Some authors, in fact, list completion under the more general heading, "short answer" or "simple recall." The length of answer, variety of possible answers, and factual or theoretical nature of the subject matter will all enter into the relative objectivity or subjectivity of scoring. Most statements with an omission or two can be changed around into the form of a direct question calling for a brief answer. Either form is good usage if the meaning is clear, but perhaps the direct question can best be left to a later chapter. Many standardized tests of achievement and aptitude have employed completion forms of one sort or another with successful results.

In making suggestions and rules for writing in this form, first consideration should be given to making the purpose of the question clear to anyone having the background of preparation to answer it. Many of the rules for true-false and multiple-choice items will be useful for this purpose, but a few special ones may be helpful for this group particularly.

1. *The statement should indicate the type of response desired.* This does not mean that clues should give away the answer, but rather that the statement should designate whether a city, state, or country is to be named, a date is to be given, or a technical term is to be supplied, etc. If this is not done, the variety of possible acceptable answers will sometimes be almost infinite, and grading therefore next to impossible.

VIOLATION: Experimental psychology began in _____.

Possible answers to the above include *Leipzig, 1879, and Germany*. Credit would have to be given for any of these in order to be fair, even though the date may have been the information wanted.

Frustrating confusion could have been avoided entirely here by specifying:

IMPROVED: Experimental psychology began in the year _____.

or

The country in which experimental psychology had its beginning was _____.

Specifying what kind of technical term is to be supplied may be essential, as in the following example.

VIOLATION: A person who deliberately sets fire to houses or buildings is called _____.

Possible answers include the popular designation *firebug*, the legal term *arsonist*, and the psychological term *pyromaniac*. Probably the general subject-matter area of the examination would not be sufficient to make clear which of these was wanted. Therefore none of them could be counted wrong in scoring.

IMPROVED: The legal term for a person who, etc.

Perhaps the worst offenses against this rule are verbatim quotations from a textbook in which so much has been omitted that originality may be the only possible source of a sensible response. The different quality of answers thus obtained will constitute a problem similar to essay scoring at its worst.

VIOLATION: Intelligence is _____ the _____
to _____.

The author of this item was asking his students to reproduce a rather obscure statement from their textbook to the effect that Terman once said that intelligence is essentially the ability to do abstract thinking. Obviously not enough was given to show the best students what was wanted, since the answers obtained included everything imaginable except the quotation. Examples are: Intelligence is *considered* the ability to adapt to new situations. Intelligence is *correlated with* the ability to make good grades in school. Intelligence *has got to be above* the average to succeed in college work. Intelligence is *regarded as* the foundation to our social order. Intelligence

is *what they call* the *division of the army assigned to find out things about actions of the enemy*.

It is not clear from the form of the sentence that a definition is wanted, but the first student in the above paragraph gave an acceptable one that should obviously receive full credit. The second answer is a true statement of an idea, not quoted verbatim, contained in the textbook, and should probably be given full credit. Answers of increasingly inferior quality follow these two. The last is a desperate attempt to use common sense and perhaps practical experience where either the subject matter was completely lacking or the student could not fit his fragmentary ideas into this skeleton. Fairness and objectivity in grading can be achieved by being specific, as follows:

IMPROVED: Terman said that intelligence is essentially the ability _____

2. *Clues that help too much or actually give away the answer should be avoided.* To include this principle here seems at first mere repetition of what was already said in regard to other forms of questions, but some special illustrations deserve attention in connection with completions. Simple as they may seem, clues of grammatical construction often escape notice. The following is taken from a rather out-of-date examination in psychology, but it illustrates this point well:

VIOLATION: A person born without much intelligence is designated as an _____, while one who deteriorates or loses what he has in the way of mind is termed a _____.

The first blank calls for *ament*, while the second calls for *dement*. If both terms had been encountered in study, but not thoroughly learned, the student might confuse the two. If this is true, the articles *an* and *a* will enable him to insert the right one of this pair in the right place, even though he might have reversed them without this clue. Omission of these clues would make the question suitable for the content upon which it is based.

Clues may sometimes be found in the body of the material given. The following seems rather obvious:

VIOLATION: An adolescent boy has conflicting emotions because he wants to conform to his age group, but if he does he will displease his parents. He is suffering from _____.

The answer is *emotional conflict*, according to the teacher who used this item, and both words in slightly modified form have already been given. A better wording might be:

IMPROVED: An adolescent boy feels disturbed because he wants, etc.

Even the length of blanks may eliminate some wrong answers and suggest right ones.

VIOLATION: The state having the largest area is _____.

California, New Mexico, and other states with long or medium-length names are eliminated by the shortness of the answer blank. The student may still have to decide among Utah, Idaho, and Texas, but his chances of guessing Texas would be greatly reduced by making the blank longer. Long blanks make for greater legibility even though they may use up more paper than needed.

3. *Make a liberal scoring key with partial credit allowed if necessary.* Although it might be desirable to have one and only one correct word or phrase for each blank, this ideal is seldom practical in some types of subject matter. The student who has the knowledge to classify something correctly in a broad category deserves some credit, even if he does not earn as much as the one who can place the same thing in a more specific classification.

VIOLATION: An IQ of 40 classifies a person as _____.

Key answer: *imbecile*

Clerical help employed to grade papers would follow this key and give no credit for *feeble-minded*, *mentally deficient*, or *severely retarded intellectually*. Assuming that the key answer is really the best one, it might receive twice the credit allowed for the three possibilities mentioned above, but there is no basis for giving these three more general answers zero. They do not represent ignorance of classification as to ability, but merely a less exact knowledge. Expert checking of grading may often be necessary where completions have been used. An all-inclusive key may be almost impossible to

make on some items. A number of examples have come to the writer's attention in which a student deserved full credit for a highly original and entirely correct insertion that his professor had not recognized as possible when he wrote the question.

Matching Items. In material of this type there are two columns of items. Those on the left are to be paired with those on the right in a manner designated in the instructions. Dates may be matched with events, names with events, authors with books, terms with definitions, rules with examples, and so on. Nonverbal illustrations include tools with their uses (in pictures). A large number of learning situations require that associations of one sort or another be made between various ideas. It is in these situations that matching test items become most valuable when properly constructed.

The chief disadvantage of matching is that it is not very well adapted to the measurement of real understanding as distinguished from rote memory. Perhaps occasional exceptions to this statement can be found, but most examples of this form call for rather simple factual association. Complex problems of interpretation of data or facts would hardly fit the matching form very well. A second disadvantage is the great likelihood that irrelevant clues will be present. If the examinee knows several of the correct pairs, he may be able to answer several others successfully by a negative process of elimination of highly improbable responses until the situation is reduced to one or two possibilities. He may not be sure at all of either of these, but he has one chance in two of guessing right. In some cases he may find by this negative approach one and only one possible answer, even though he never learned the correct association at all. A still further disadvantage is the amount of time required to answer matching material, especially if there are 15 or 20 pairs in the group.

In formulating rules for construction of this type of material, emphasis will be placed first upon the selection of the material itself. The rules on this point are not in themselves sufficient to ensure that the items will be good. Only experience in working with them and having them reviewed by someone else to catch various weaknesses overlooked by the writer will finally result in the production of really valid material. Some teachers and other workers in the field of test development are apparently unable to detect the clues readily discovered by the clever examinee until a reviewer points out these

devices. Only by experience can the worker acquire skill in this type of detective reviewing of his own productions.

Secondly, the rules will cover helpful suggestions regarding the format. There are a number of variations of instructions in common use today. A few will be illustrated here.

1. *Indicate clearly on what basis matching is to be done, and follow this pattern throughout the construction of each exercise.* If dates and events are mixed up with names of places and events in the same group, there will be clues that encourage a process of elimination or a negative approach. Similarly, definitions and terms should not occur in the same group with rules and examples. Directions and column headings should both indicate the basis of matching.

VIOLATION: Place just one letter to the left of each number.

- | | |
|----------------------------|-------------------------------|
| _____1. Terman and Merrill | A. Adult intelligence test |
| _____2. Aptitude | B. Revised Stanford-Binet |
| _____3. 1916 | C. Innate capacity to learn |
| _____4. Validity | D. Correlation with criterion |
| _____5. Wechsler | E. Terman revision of Binet |

This example violates several rules, as will be seen later, but it is not unlike many less extreme violations frequently seen. The kinds of pairs include author-test, term-definition, and date-test. This lack of any common principle of association makes a negative approach more fruitful than it would be otherwise, and thus encourages a search for clues, which are present in this case. The examinee might, for instance, think out loud about as follows: "Well, Terman and Merrill—that could be Terman revision of the Binet. No, I remember that's 1916. I'll put E for 3 now. Well, that lets that one out, so it must be B for 1, I guess. All right, aptitude, that's either C or D, because they are the only definitions there. I don't know, but let's see. There's another term, validity; that's a correlation of something. I remember reading that. I'll put D for that, and that leaves C as about the only thing that would fit 2. Wechsler—Where've I heard of him? Must be A, that's all I've got left."

All answers contained in this lingo are correct, but the approach is negative. A clever individual with a few hazy ideas about this particular subject has managed to fool the test. His score is perfect, but his thinking is cloudy, because he did not study this material

systematically in the first place. The examiner would therefore do well to ask himself what the simplest mental process would be that would bring success on this exercise.

2. *Check each group for clues.* This rule may avoid the probability of such a procedure as that quoted above. Rule 1 may have been followed, but clues may still be there. A number of these have already been discussed under other forms of items.

3. *Discourage a process of elimination by including more responses than are needed.* Increasing the right-hand column in our example by two items that would not fit anything on the left, but that would be plausible distractors (as in multiple-choice), would increase the difficulty of the group and catch our vague, poorly informed thinker. These distractors should be between good items on the list, not down at the bottom of the column.

4. *A list of short items will be easier to work with if they follow some logical order.* Dates can be chronological, authors or terms alphabetical, etc. Following this rule will shorten working time.

5. *The length of an exercise or group of pairs should be about 5 to 15.* Too few pairs encourage guessing, while too many increase working time.

6. *A group of pairs should be all on one page.* Leafing over pages wastes time and prevents the proper operation of the gestalt principle.

Now the revision of our original example may be undertaken. New pairs will obviously be needed to make the group homogeneous in some way. Column headings will need to be added, and the instructions will have to make clear the basis for association, or mental set, as it is sometimes called.

IMPROVED:

INSTRUCTIONS: Match the following tests with their authors. Put just one letter before each number.

<i>Authors</i>	<i>Tests</i>
C 1. Bernreuter	A. Mechanical Comprehension
E 2. Kuder	B. Vocational Interest Blank
B 3. Strong	C. Personality Inventory
F 4. Terman	D. Thematic Apperception
G 5. Wechsler	E. Preference Record
	F. Revision of the Binet-Simon
	G. Intelligence Scale for Children

Items A and D at the right are distractors. They do not fit. Item 5 is a difficult one, since G is a recent test associated with this name and not the one more commonly known. The authors are alphabetical. The instructions are usually clear in this form if those taking the test have ever done matching before. If not, then an example might have to be given.

Grading of matching should be completely objective. The form shown in our improved example will make for faster and more accurate grading than lines drawn connecting the pairs. In making the key, care must be taken to recognize two possible correct answers for one item, should such be entirely reasonable. Complete avoidance of more than one answer may not be possible or even desirable. Many subject-matter fields lend themselves readily to this form of examination question, but it is probably less useful as a tool than multiple-choice form. Complete examinations made up exclusively of matching are likely to be much too narrow in scope. This excellent technique must be used sparingly and wisely.

Rare Test Forms. In various stages of experimental development are a number of interesting new types of test items. Some of these have so far proved to be of value only in limited areas of subject matter, while others remain in an embryonic stage in which their usefulness in any field has yet to be demonstrated. A few of these will fit best into the discussion in the next chapter. Some of them have no place in this volume. Worthy of brief comment here, however, are attempts to find some objective way to evaluate abilities which many authorities believed could be judged only subjectively in essay tests. One of these is the ability to organize facts, giving ideas in some kind of a logical sequence. An item intended to measure this important trait in human thinking may take the form of an arrangement question. The following is an example taken from the writer's lecture notes on thinking. The material from several sources seemed to show some general agreement upon seven stages in scientific discovery and invention.

EXAMPLE: On the lines at the left, indicate the best arrangement of these steps in scientific discovery and invention. Letter the steps in what you think would be the best order.

- _____ A. The idea or solution appears suddenly as a whole.
- _____ B. Think of all possible solutions.

- ___ C. Perfection of the idea.
- ___ D. Perceive a need or difficulty.
- ___ E. Survey all previous work relating to the problem.
- ___ F. Analysis and clarification of the problem.
- ___ G. Weigh solutions for their advantages and disadvantages.

The best order is probably defensible from several sources as D, F, E, B, G, A, C. If this order had been given to a group of students as best in the lecture before the test, rote memory might be a source for many right or nearly right answers, but if the students had not found a prescribed order anywhere, the question would be an organization task rather than a mere rote-memory task. Then the question of several complete-credit or several partial-credit arrangements would arise. Controversy over this point might entirely invalidate the item. The difficulties of constructing an item of this sort seem almost endless. Only consultation with several authorities could establish the merits of different answers. Only those on which expert opinion was generally in agreement would be worth retaining. Not all arrangement items would belong in the same category with the one given above. Steps in mathematical problems and in certain experimental procedures in chemistry and physics, for example, are quite well fixed in sequence, since correct results cannot be obtained unless the prescribed order is followed. Such an arrangement item on an examination becomes a rote-memory item when the student is provided, through lecture or text, with a list of steps in their order. He can memorize them word for word instead of reasoning out the how and why of the sequence. A problem requiring a logical arrangement of entirely new material may become a good thought question if the key is worked out carefully. Credit must often be given for more than one possible arrangement in such cases, and the opinion of more than one expert is nearly always essential before the worth of such an item can be demonstrated.

Among the characteristics of an individual with good study habits may be mentioned the ability to distinguish between important ideas and subordinate details. The student at either high school or college level who clutters up his mind with a great mass of isolated facts that are not organized in any way is the one who often blocks on tests and is unable to recall what he has learned. The original learning

is of course at fault in this instance. Although essay questions may, as is claimed, evaluate the extent to which one has mastered this organizing skill, these may not be the most effective tools for doing so, at least not on a large scale. A new experimental kind of objective question may have merit for this purpose if a tryout can demonstrate its validity. This new tool may tentatively be named an objective outlining item. The exercise, consisting of perhaps 10 to 20 items, can be planned in this way: First, take an outline that appears subjectively to be so logical as it is arranged that any reorganization of it would probably result in an inferior product bordering on the incoherent. Not every outline in a book or elsewhere will meet these specifications if it is studied closely, but probably there are some that will. Then mix the topics up in random fashion, and, without any differentiation as to heading or indention, begin each at the left margin and number them consecutively. Finally, write instructions of the simplest sort possible requiring the examinee to designate whether each topic is a main heading, a subordinate heading, or a detail. The most difficult step will come in working out the scoring key. Here the examiner may make a serious error in relying only upon the original arrangement of the outline or upon his own judgment as to acceptable approaches. Other organizations of the material may be possible that really have equal merit, and the examinee could perhaps defend his "wrong" answer if he had a chance to explain his approach to the problem. Exercises of this sort, therefore, must be used with extreme caution until tried out and studied thoroughly.

An example which is at present still in the stage of tentative use has been constructed from the outline in an unpublished paper on personnel practices in public employment. The outline was modified somewhat, and only a portion of it was included in this exercise. The exercise has been tried out so far on only a half-dozen examiners in a public personnel agency, and the data at hand are insufficient to establish its merit definitely. It is reproduced here to stimulate further work.

INSTRUCTIONS: Here are some topics taken from an outline on employee morale. They are all in mixed order with no indication as to their relative importance. On the original outline some of these were main headings,

some were subordinate headings, and some were details. On your scratch paper, reconstruct this outline in the most logical way you can find. Then, on the line to the left of each topic below, print

 M if it is a main heading

 S if it is a subordinate heading

 D if it is a detail of less significance than a subordinate heading.

- _____ 1. A typist works harder if he knows what is done with the three carbons he is assigned to make.
- _____ 2. The principle of equal pay for equal work should be followed.
- _____ 3. A clerk who realizes the probable consequences of her errors will try to avoid them.
- _____ 4. Employees should feel that they are making a real contribution to the organization.
- _____ 5. Pay increases should be given at regular intervals if work is up to standard.
- _____ 6. The "why" of routines that appear absurd can usually be made clear.
- _____ 7. A fair pay plan must be established.
- _____ 8. A six-month interval is suggested for entrance-level jobs, while one- or even two-year intervals are suitable for higher-level positions.
- _____ 9. Each should know how his job fits into the total picture.
- _____ 10. Job-performance ratings help to decide whether employees earned raises.

The original outline and key answers are given below for detailed study of the structure of the exercise.

EMPLOYEE MORALE

- I. A fair pay plan must be established.
 - A. The principle of equal pay for equal work should be followed.
 - B. Pay increases should be given at regular intervals if work is up to standard.
 - 1. A six-month interval is suggested for entrance-level jobs, while one- or even two-year intervals are suitable for higher-level positions.
 - 2. Job-performance ratings help to decide whether employees earned raises.
- II. Employees should feel that they are making a real contribution to the organization.

- A. Each should know how his job fits into the total picture.
- B. The "why" of routines that appear absurd can usually be made clear.
 - 1. The typist works harder if he knows what is done with the three carbons he is assigned to make.
 - 2. A clerk who realizes the probable consequences of her errors will try to avoid them.

Key: 1. D, 2. S, 3. D, 4. M, 5. S, 6. S, 7. M, 8. D, 9. S, 10. D.

There is some difference, in the tryout answers so far, in the order in which topics occur in the reconstructed outline on the scratch paper. This is to be expected, but it should not have an effect upon the *classification* of the topics as main headings, subordinate headings, and details. In fact, so far the writer has found very little valid objection to the answer key given above. That such an objection is possible must be admitted until the material is further evaluated through use.

An obvious disadvantage of this objective outlining problem is the amount of time required to answer a complex group of items of this sort. An essay question would probably not lengthen the working time. Yet the scoring can probably ultimately be defended better than that of an essay question if sufficient research has been done beforehand. Hours spent in this research will be saved in the long run if the volume of scoring is considerable, since checking the letters from a key can be quite rapid.

In this chapter, the broad scope of objective types of test items has been demonstrated. The advantages and disadvantages of each form have been reviewed, and rules, where well established, have been stated. Objections to objective tests have been shown to be based in many instances upon examples constructed by persons lacking skill in test construction who could not see possibilities beyond rote memory in these types of examination material. Objective tests have some limitations that are valid, and these will be discussed in a later chapter.

SPECIAL PROBLEMS IN
OBJECTIVE-TEST CONSTRUCTION

THERE are a few areas of subject matter in which the conventional forms of test items such as true-false or multiple-choice may not fit so well as would some modifications of these. In this chapter, therefore, a few of these modified procedures will be analyzed. The first section of this chapter reports research conducted by the writer, which has its applications in an academic situation, high school or college, as well as in public and private personnel work.

An English-usage Test for Clerical Workers. Many different forms of question have been devised to measure the ability to spell, punctuate, use correct grammar, and employ the right word in the right context. The mechanics of English are important to the high school, business college, and college or university student as well as to employees in many different kinds of job, especially clerical, typing, and stenographic. One classical academic form of test in this field is the straight dictation by the examiner of sentences, which are taken down in longhand. Grading of such papers is laborious. Printed sentences in which errors are to be corrected in pencil represent some improvement on the dictation method so far as scoring is concerned, but locating the right and wrong responses and counting them is still rather tedious. Multiple-choice items which isolate within a sentence some particular problem of punctuation, spelling, or grammar are easily scored, but they are usually not so difficult as sentences in which the errors are not made obvious by selecting some word or phrase for three, four, or five choices. Some civil service examinations are now given in which sentences appear with four numbers scattered along above each, with one number directly over the error as shown in sample A below.

1 2 3 4

SAMPLE A: The man who I wanted to see is occupied this afternoon.

In this illustration, as well as in most other items of this form which the writer has observed, the error is made too obvious by its position directly under a number. The incorrect numbers are seldom above anything that might wrongly be considered an error. Although most of the material in these tests contains excellent illustrations of the most important principles of English usage, the sentences as presented are a considerable departure from the situation usually encountered in an office where letters, reports, and other material would have to be prepared with the utmost care, proofed for errors, and corrected.

In order to measure proofreading skill such as would be essential for jobs in these classes, and in order to increase the difficulty of recognizing errors while at the same time retaining the ease of scoring characteristic of the multiple-choice form, the writer originated a new manner of presentation of English-usage material. An elaborate system of symbols designating grammar as G, usage as U, etc., was felt to be superfluous because it would involve following complex directions, thus introducing an additional difficulty factor which was not intended to be measured in this section of the test. Also, it seemed that the duties of clerical positions would not require the candidate to define the type of error in this way. Even though it might be ideal for him to recognize the principle involved, he needs only to sense where something is wrong in order to identify the mistake by number. If, in an academic situation, the aims included labels for the different kinds of mistake found, then an identifying symbol could be called for in addition to the number.

In the system of presentation used in this particular research project, each sentence is divided into four sections by means of diagonal lines. Each section is numbered, but the number does not necessarily fall directly above the word, phrase, or punctuation mark that should be corrected. Some of the sentences are entirely right and are to be answered by *R* instead of by a number. The answer to an item is always either a number or *R*. Sample B shows the manner of presentation in this research.

1 2 3 4

SAMPLE B: The man who/ I wanted to see/ is occupied/ this afternoon.

In this illustration, there are no clues as to the location of the error, and it is therefore more difficult than sample A. Most of the sentences are complex enough as to vocabulary or structure so that a candidate who is uncertain about any of the accepted rules might easily think he had found an error in spelling, punctuation, or word usage in a section that is actually entirely correct.

At the time that this research was in progress,¹ the Louisiana Department of State Civil Service, where the writer was employed as consulting psychologist, was giving a basic clerical test for entrance-level positions, including a clerical-aptitude test of the number- and name-matching variety, some practical office problems in arithmetic, following directions, and this English-usage test. There were 15 of the sentences to be proofed, and they covered a fair variety of errors commonly found in office material in several agencies. Four equivalent forms of the entire battery were ultimately developed, since these clerical tests were often repeated at least four times a year. Later there was a need for more difficult items of this sort for higher-level positions, such as office supervisors and personnel assistants.

A statistical procedure to be described later in this volume was then applied to the results of a tryout on a large group. Although the item-test correlations were not quite so high for the English-usage items as for following-direction and office-arithmetic problems, the sentences held up fairly well. Their difficulty level varied greatly, and there were some that turned out quite unsatisfactory from a statistical standpoint, either because they were too easy or because somehow they did not correlate at all with what the test as a whole had measured.

Other criteria need to be considered besides item-analysis data to determine whether a sentence is fit for retention. Experts were selected who were at the time teaching business-letter writing and other related subjects at Louisiana State University. These specialists were asked to review all 120 items then in use to determine whether every principle illustrated was defensible in terms of modern practice. Several were found obsolete. Rules, particularly with regard to punctuation, have undergone some change. Those who went to

¹ Bean, Kenneth L., "The Development of an English Usage Test for Clerks, Typists, and Stenographers." *Educ. Psychol. Measmt.*, 1946, 6:331-339.

school before 1925 were usually taught that commas must be used under some circumstances where today it would be acceptable to omit them. Where differences of opinion between present and past authorities were found, the sentence was changed or omitted, since the error would not be clearly defensible as wrong.

On the whole, the item analysis revealed that the most difficult sentences were those involving a choice between *who* and *whom*. Punctuation ranked second in difficulty among the problems presented by this particular group of sentences. Then followed, in order, word usage, spelling, and capitalization. The above statements should not be taken as generalizations, since a very small sample of each type of error could be included in a test section of this length. Another group of such sentences would probably change this order somewhat.

In 1944, the 120 items then in use were given to 200 college freshmen at Louisiana State University. The scores were then correlated with the Purdue Placement Test in English. The Pearson product-moment r found was .71 with PE .017. The reliability (split half) was found to be .84. Therefore, there is some reason to believe that this new form of item has merit.

Another kind of item seems to be needed at higher levels in which employees compose letters of something more than routine nature. The writer has for some time used multiple-choice form for judgments as to which is the best of several phrases or sentences to achieve a given purpose.

EXAMPLE: Which of the following is best to begin a business letter?

1. Your letter of Dec. 2 received and contents duly noted.
2. Yours of December 2 received, we beg to reply that we have shipped your order.
3. Thank you for your letter of December 2.
4. Your order of Dec. 2 received and shipped same today.

The problem of usage here is one that involves the whole opening sentence. Therefore the multiple-choice form fits the problem better. Choice 3 is a direct and simple statement that contains no cliché or overused expression. Presumably it will be followed by a statement to the effect that the order was shipped.

Employment Tests for Illiterates and Near Illiterates. In dealing

with written test material for janitors, hospital attendants, bricklayers, painters, and others in this broad grouping, great care must be taken to avoid difficult vocabulary except for the terms of the occupation with which they probably have very frequent contact. A somewhat modified test must be employed. Very simple sentence structure is absolutely essential. An illustration of this point was given in the discussion of appropriate wording of multiple-choice questions in job language (Chapter 3). Often the yes-no form is better than the multiple-choice, in spite of the guessing factor. The concepts of true and false are too abstract for individuals with limited educational background. Many jobs require a minimum of verbal comprehension. Emphasis upon verbal intelligence or general information of an academic nature in tests for these jobs will lower their validity. Ability to follow instructions and knowledge of routine procedures on the job itself should probably be the two principal aims of the written part of the personnel-selection process. A supplementary performance test will nearly always be essential.

Consideration of the performance or work-sample test will be deferred until a later chapter. At this point a few suggestions for the preparation of the written part will be discussed. To secure the best cooperation and the highest degree of interest in the material, the items must relate to situations that are as practical as possible. Usually instructions should be read aloud by the examiner. As little written matter as possible should be given to the candidates. Marking of answers should be done in the simplest manner possible, and small groups rather than large ones will make possible some individual help where needed on the method of marking the answers. An academic kind of testing situation actually frightens some very capable workers in manual occupations and causes them to make foolish mistakes that come about through confusion rather than through any lack of aptitude or knowledge.

In the experience of the writer, this emotional blocking has often resulted in a lowered rating or even occasionally in a failure on an employment test. One exceptionally efficient foreman on the docks of an important port on the Mississippi River had been working on a provisional appointment for eight months or more during World War II. The time came for him to take a written test for the purpose of gaining full civil service status in his job. The state agency for

which he had worked reported that he had displayed exceptional ability to carry responsibility for supervision of the distribution of unloaded material from ships to warehouses and to places where it was loaded on freight cars. He could remember numerous details about this work and made very few errors. The written test which he took was yes-no and multiple-choice form with simple sentence structure and vocabulary. Except for daily encountered shipping terms, the words were those any fifth-grader could understand. The employee stated that he had quit school in the eighth grade, over fifteen years before. The test situation made him very nervous. He squirmed in his chair, and he was unable to add a simple column of figures correctly. (He did longer problems of the same sort every day with seldom an error.) He made foolish mistakes in failing to follow instructions. The rating which resulted was a failure. A second try at the test several months later was no better, but the agency could not find among those who passed the test one who could learn the job effectively. This is one of the exceptions to the rule that is often chosen by those critics of testing who are trying to prove that scientific personnel selection is no good. Few cases are that extreme, but most men and women who have been out of school for a while dread any kind of a test if their jobs depend upon it. They need considerable reassurance in order to do their best. The situation must therefore be as much like the daily work to which they are accustomed as possible.

The above case does not support an argument for a low standard of performance for applicants in the types of occupation in question. Rationalizing failures on the basis of nervousness can be carried too far. Most people who "go to pieces" on a test will be likely also to break down in their performance if placed under any stress on the job. The caution proposed here applies to the kind of material encountered on the test. A man who knows carpentry and is confident that he is a skilled tradesman will not ordinarily become frightened or confused if asked to solve the kind of problem or to do the kind of task he encounters regularly in his work. He will be filled with apprehension and bitter resentment if confronted with a textbook problem in arithmetic formulated in theoretical language and having no bearing whatever so far as he can see upon being a successful carpenter.

Keeping these general precautions in mind, the test constructor can work best from his observations of the job, the specifications, and such manuals or books as may be a little clearer in their explanations than would be most of the workers themselves. An example of a practical test problem can be followed through here to illustrate one procedure found by the writer to be quite effective. The position was highway maintenance foreman. Duties included supervision of a small group of men, three to a dozen, perhaps, depending upon the work to be done, in such jobs as grading, repairing concrete highways, cutting weeds near the road, and repairing bridges. The average educational achievement of men performing several of these jobs successfully at the time was about sixth grade. Experience in construction work was required. Ability to follow somewhat detailed oral instructions seemed essential. The comprehension of written material beyond reading of road signs and other very simple matter was not necessary.

Instructions for Tests. The problem prepared for the first of several sections of the written test was a set of instructions in modified form very similar to those actually given one morning in the highway department shop at a large center. The modifications made in these instructions were largely omissions of details that would give decided advantage to the applicant from the particular locality in question. Answer sheets were distributed which had only question numbers and the words *yes* and *no* after each number. The examiner had a set of instructions which he read aloud. He also talked informally to the men to motivate them as well as possible. The instructions are given below:

To Be Read by Examiner: Listen carefully. Let us suppose that I am telling you about a job I want you to do. Try to remember as much as you can of what I tell you, because I am going to ask you questions about it. Are you ready now? This morning I want you to drive that truck out the road that goes south from the university. Take three men with you. Pick up a couple of weed cutters and a power mower from the tool shed. When you get out about a mile and a half, you'll come to a filling station on the left-hand side. Start cutting the weeds at the right side of the road right at that filling station and keep going until you come to a gravel driveway just before you get to a grocery store on that same side. Then check the bridge about a quarter of a mile farther down the road to see if it needs

any fixing up. I put some tools in the truck already that you might need to fix the bridge.

Now I will read question 1. Did I tell you to take the road going west from the university? If the answer to that is *yes*, draw a circle around the word *yes* after number 1 on the sheet you have. If the answer is *no*, just put a circle around the word *no* after number 1. Be sure to put the circle after the right number each time and around *yes* or *no*. Does everyone understand? (*Pause.*) Question 2: Are you going to take three men with you? Put a circle around either *yes* or *no* after 2. (*Pause.*) Question 3: Are the weeds to be cut on the left side of the road? Answer question 3 now. Question 4: Are the weed cutters already in the truck? (*Pause.*) Question 5: Is the filling station on the right side of the road?

There are 10 questions in the entire exercise. Then another job is treated in the same manner. Later single questions are asked about the use of common tools. These are in yes-no form. Picture matching would have done just as well.

A question may arise about the use of somewhat ungrammatical expressions in the instructions, such as "get to a grocery store" and "if it needs any fixing up." Actually, the instructions overheard were considerably edited in arriving at the version quoted here, and the writer felt that any further refinement of style would result in an unnatural literary flavor that would never occur in the actual working situation. Thus a compromise between the language actually observed and strict conformity to conventional usage was felt to be the best for this particular group.

Oral presentation seemed to be quite suitable here, and often in the writer's experience it has worked well for simple multiple-choice forms. Lengthy wording nearly always makes a second reading necessary, and often it leads to complete confusion. This observed fact does not mean that men in these occupations are unable to understand anything beyond the very simple. Automotive mechanics and other skilled tradesmen often deal with very complex material, but it is nonverbal material. The academic argument that anyone who understands something can explain it does not seem to be supported by practical experience. If a mechanic can, without preliminary trial and error, locate the part of a machine that is not functioning properly and correct its functioning with a series of complex operations, he obviously understands how the machine works.

Frequently he cannot tell someone else in any coherent fashion how he did it. He may not be able to tell why the changes he made resulted in the machine operating smoothly. If he does not have to teach anyone else his trade, he does not need to explain. Mechanical aptitude, not verbal ability, will then be essential for success on his job. Ideas that are not formulated in words but in visual images, perhaps, can probably be said to constitute a form of understanding. If success in repairing a machine results rather accidentally from a series of trial-and-error operations with no insight, this form of non-verbal understanding would be lacking, and qualification for a mechanical job would be questionable.

Measurement of this nonverbal understanding, if such terminology may be allowed, is important for many manual occupations. Lack of verbal facility, either in comprehension or in expression, should not penalize a candidate who can do what the job really requires. Therefore the written test must measure nonverbal aptitudes largely, and it must usually be supplemented by a good work-sample test. Since workers in test-development research are usually very superior in verbal aptitudes and skills, it is difficult for them to realize that verbal inferiority is not a handicap in many occupations. They also find that much practice is required in order to simplify adequately the language of the examination.

Reading Comprehension. Much work remains to be done in the measurement of reading skills, particularly at college and adult levels. Existing tests have many inadequacies. In high school, in college, and in many types of practical work, reading is important. There seem to be, in fact, so many types of reading disabilities that no two cases are found to be exactly alike. Many students reach college with inadequate study habits. Testing for verbatim memory is not enough. Comprehension amounts to converting words into ideas. Readers can perceive and perhaps even mumble words which they can later reproduce like a phonograph and still grasp no meaning. Meaning implies relationships between one idea and another, either within the material or between material from other sources and ideas in what is read. These associations, which give what we call meaning to what we read, occur within one sentence. They also link together what is contained in a paragraph, a chapter, a book, and indeed all cultural experience accumulated by an individual.

Differences among readers, then, can be described in terms of the number, breadth, and extent of these associations. Some accumulate small isolated units only. Each sentence is separate, not connected with any other. Some readers can go as far as the paragraph. Still others have a fair grasp of single chapters. The best readers are very rich in the associations which add to their enjoyment and use of what they read. The sources from which meaning may come are almost infinite.

A few forms that are useful in evaluating the richness of this experience of meaning may be illustrated here. One that is quite common is a paragraph or series of paragraphs followed by one or more multiple-choice or true-false questions. The multiple-choice are decidedly superior for this purpose, since the true-false are apt to call for rote memory only. The items may require mere factual recall or recognition, but to be entirely satisfactory, they should, in some cases at least, call for generalization. To design them in such a way that a grasp of the central thought of the paragraph is essential in order to arrive at the correct answer requires special skill that can be acquired only by persons with verbal facility after some practice.

An example is given for study here that is based upon material on the functions of a high school counselor.

If a poorly adjusted person is told what to do to better his condition, he may or may not follow the advice given him. In either case he is probably not much benefited until he can make an intelligent decision regarding plans for himself. This decision should be based upon some understanding of himself through results of tests or through other information which the counselor may help him to obtain. When he has made up his own mind what he needs to do, what vocation best fits his interests and abilities, how he can succeed better in making friends, etc., he will carry out his plan with more motivation, more persistence, and clearer understanding than he would ever have done if someone else told him what to do, no matter how much respect he might have for his counselor's opinion.

1. The above paragraph means most nearly that
 - A. the best counselors often give advice freely
 - B. a good counselor helps his client to make plans for himself
 - C. a counselor can be little if any help to a person, since that individual must make his own decisions independently
 - D. a good counselor tells a person the best vocation for him to choose.

2. Ordinarily the client carries out plans with the most enthusiasm, according to the above paragraph, if
- A. the counselor makes all important decisions
 - B. he has great respect for the opinions of his counselor
 - C. they are primarily the product of his own thinking
 - D. he disregards entirely any information given him by a counselor.
- C

The first question has in the premise the expression "means most nearly" primarily to answer the possible objection that choice B, the best answer, does not really cover all of the content of the paragraph. It obviously hits the central idea, while the other choices are contradictory to the thought of the material. The item could have been made more difficult by including among the wrong choices some decidedly subordinate but true detail of the content, such as:

Results of tests often help a person understand himself better.

The above is not a general thought from the material as a whole, and could not be called the best answer, yet a reader who does not assign ideas to some scheme as to their relative importance as he goes along through a paragraph might choose this answer as perfectly satisfactory. An individual who fails question 1 as it stands has missed the thought entirely.

The second question is closely related to the first, and an examinee who answers the first one right is very likely to know the correct answer to the second one also, but the inclusion of both questions reduces the likelihood of chance success and adds a little to the thoroughness of measurement of comprehension.

The same general pattern just described has found wide acceptance in personnel-selection tests where the accurate interpretation of legal material is required. The writer has made extensive use of this form for public welfare administrators, juvenile-probation officers, and office managers in governmental agencies primarily concerned with administering a law or system of prescribed regulations. The validity of questions on legal material is very difficult to establish, however. Law is a controversial field in many areas, and thorough review by more than one qualified authority will be necessary in most instances.

Still another valuable application of the above pattern is found in the interpretation of scientific laws and principles. The question may call for a recognition of a restatement of the same principle in the premise and in the correct choice, or it may set forth the law in the premise and give an example in the correct answer. Inaccurate versions or examples that would not follow the rule make up the incorrect choices.

SAMPLE: "Equal relative differences are equally perceptible" means that

1. the sensation increases in arithmetic ratio when the stimulus increases in geometric ratio
2. if the subject can barely tell the difference between 50 grams and 52 grams, then he can just barely discriminate 100 and 104 grams.
3. Both of the above are correct.
4. Neither of the above is correct.

3

A person who can restate and apply the Weber-Fechner law has thoroughly comprehended it. He must be able to do both to recognize choice 3 as the *best* answer to this question.

Another pattern that will carry the reading-comprehension test a little beyond the usual degree of thoroughness is one that is in a rather experimental stage at the present time, although it can be found in a few tests that have been tried out on large groups. For lack of any better designation, it might be called the true-false-zero item. This item consists of a basic paragraph followed by several statements, each of which is to be marked as follows: *true* if it is fully supported by the content of the paragraph or if it is an entirely reasonable inference from the material given in the paragraph; *false* if it is definitely contradicted by the paragraph; or *zero* if it is a statement on which the paragraph gives insufficient evidence either to support it or to contradict it. The instructions must stress that only the material given, and not common sense or other sources, must be used as a basis for answers. The instructions are obviously intended for a rather advanced level of reading skill and thinking ability. The true-false-zero item is rather unsuitable for lower grades in the schools and for employment tests for most entrance-level positions. For technical and administrative jobs it probably has considerable value, but only further tryout can demonstrate its worth.

Failure to grasp ideas that are in the material read is one common fault, but quite common also is the habit of going beyond what is actually there much further than is justified. Projecting one's own ideas into a paragraph to the extent of believing that the material actually contains those ideas will lead to errors in thinking and lack of objectivity in evaluating what is read. The true-false-zero item is perhaps one rather objective method of detecting subjectivity. It deserves further study.

An example will perhaps be helpful in exploring some of its possibilities.

Reading disabilities may result from a wide variety of factors which may operate separately or in combination. No two cases seem to be exactly identical. Often the inability to read is a symptom of severe tension in the child's home situation which, when relieved and replaced by satisfactory home adjustment, will result in prompt, spontaneous recovery without remedial reading training. Very slow reading most often results from poor perceptual habits. The pupil almost literally spells out his words, taking in a very narrow span with each fixation of the eyes.

DIRECTIONS: In the left margin before each of the following statements you will find T F O. If the statement is fully supported by the content of the paragraph or if it is an entirely reasonable inference from the material given in the paragraph, draw a circle around T to show that it is true. If it is definitely contradicted by the paragraph, mark it false by encircling the F. If it is a statement on which the paragraph gives insufficient evidence to support or contradict it, encircle the O. *Do not* use other sources of information for answers.

- | | | | |
|-----|-----|-----|--|
| T | (F) | O | 1. One cause can usually be found to explain reading disabilities. |
| T | F | (O) | 2. Reading disabilities are like speech defects in that both may be symptoms of severe maladjustment in the personality. |
| (T) | F | O | 3. Relief of tensions at home may be all that is needed to cure a reading difficulty. |
| T | F | (O) | 4. Very quick exposures of words on cards may result in a wider span of perception. |

The first sentence of the paragraph contradicts question 1, which has therefore been marked *false*. Question 2 is probably a true statement on the basis of actual clinical experience, but it is not verified by the material given. It does not necessarily follow from what is

given that speech defects are also a symptom. Therefore the *zero* is the only defensible answer. Question 3 does not depart from the given content enough to leave any clear thinker in doubt that it is to be marked T. Question 4 is also true, but one could learn this fact only from other sources, not from what is given. Therefore the 0 has been encircled.

With a combination of the above suggestions, it should be possible to do a rather diagnostic study of an individual's reading. Some students who are quite bright have been observed to use common sense in answering questions based on reading material which they did not really understand. With many test items now in use, both standardized and teacher-made, a degree of success would be possible by this method. The true-false-zero item will detect this kind of inadequacy better than any other form now in use that can be objectively scored. Its present limited use, however, would suggest that the formulation of definite rules for its construction be deferred until further research is done.

A few specialized problems in test construction can be found in measuring the ability to use graphs and charts. The multiple-choice item usually fits best for the interpretation of maps, statistical tables, bar graphs, etc., but sometimes a special form of coding item is useful, particularly for tabulated data. Memory for a problem and ability to use data from a table in solving it involve reading skills of a special type, a good memory for figures, and logical reasoning ability. Look at the example given below:

INSTRUCTIONS: The table below gives some information about several employees. Some problems follow. The answers to these problems are to be found in the table. Each problem calls for the *code number* of an employee, which you will find in the first column.

Code	Position	Salary per month	Age	Years served
1.	Office manager	\$320	46	3
2.	Typist clerk	180	22	1
3.	File clerk	170	27	1
4.	Typist clerk	190	31	1
5.	Office manager	320	57	9
6.	File clerk	170	24	3
7.	Typist clerk	180	39	1
8.	File clerk	165	21	2
9.	Messenger	135	18	1
10.	Messenger	140	19	1

— *Sample problem:* The office manager who had the longest period of service.

The code number of the office manager with 9 years of service is 5. Therefore, this number should be entered on the line in the left margin. Work the problems below.

Problems:

- 8 1. The employee who is a typist clerk or file clerk and who is not less than 19 or more than 21 years old.
- 6 2. The employee who has served more than 1 but less than 9 years, who is between 20 and 30 years of age, and whose salary is more than \$165 but less than \$190 per month.
- 1 3. The employee who is not a typist clerk, who has served 3 years or less, who earns more than \$180 per month, and who is less than 55 years old.

In the exercise, an attempt is made to formulate practical problems in finding in tabulated data the case that fits a given set of requirements. The requirements must be understood when read and retained long enough to be used as a total pattern. An attempt to take one step at a time in solving the more complex problems often meets with failure, because a case that fits the specifications in one column does not necessarily fit the requirements based upon figures in another column. Forgetting any one limitation placed upon salary, age, title, or length of employment may result in a wrong choice. Exact following of the instructions as phrased may be essential, as in the case of such limitations as "more than —" as distinguished from the quite different formulation, "— or more."

The data may cover 20 or even 25 code numbers instead of the 10 given here, and the number of columns may be increased to perhaps 6 or 7. The process of elimination may then be made a much more complex one in which many choices are possible until all conditions have been met. It is not usually good practice to call for more than one answer to each problem.

This following-directions exercise in interpretation of tabulated material is a useful pattern to follow for a portion of the test for selecting office clerks, statistical clerks, and research workers in a number of fields. The table of figures may contain population increases and decreases for a group of cities and towns for several periods of years, test scores, grades in academic subjects, and ages, or

the results of surveys in sociology, public welfare, education, etc. In any area of knowledge, the problems will follow essentially the same pattern, except that the code numbers may represent states, cities, pupils, or some other units of information. The examinee should be required to skip around at random for his solution to each problem. The order in which columns are referred to in the problem should not be uniform, but irregular and variable from problem to problem. This suggestion for construction ensures that the exercise will measure a specialized form of reading skill.

Vocabulary. Probably the best single subtest as a rough indicator of general intelligence is vocabulary. The Wechsler-Bellevue Intelligence Scale and the Revised Stanford-Binet both contain extensive vocabulary subtests. These have proved useful at all age levels from preschool through adult. These two individual tests require the subject to state a definition. Obviously this brings some subjectivity into scoring, yet the standards for grading on both tests are remarkably clear and specific in most instances, and a reasonable degree of objectivity is attained by the careful examiner with sufficient training and experience.

The use of vocabulary on an intelligence test assumes that environment has been reasonably near normal. Vocabulary is acquired by an intelligent individual with normal or reasonably typical background for his particular culture. Exposure of the mentally defective child to the very best of cultural opportunities and the most pedagogical methods of instruction will not result in anything more than perhaps an ability to pronounce some big words. He still cannot explain what they mean except perhaps by parroting what his teacher or parent has told him. Native intelligence is essential to the acquisition of vocabulary. Yet it is not the only essential factor. Natively superior individuals from an environment where poverty and ignorance dominate the picture do not grow up normally as members of their society or products of their culture. Where books are not available to the child, and where the daily conversation of his parents and other associates is extremely restricted in content, measurement of vocabulary will result in a rating that is too low. This rating will not represent native capacity to learn and think, but it will simply show the present level of functioning, which has been adversely affected by environmental restrictions. Improvement of

the environment will, in a few years perhaps, result in a much higher score and one that is more representative of native capacity.

An analogy may be found in the physiological field. If a growing boy is confined to a small space and discouraged from exercising his muscles, he will not develop strength up to the limits set by his hereditary. He may not be endowed with the capacity to develop into a football star, but his development will be greater, up to a point defined by heredity, if he is given encouragement and training.

Thus general vocabulary usually is a valid measure of intelligence provided that a normally stimulating environment is present. One other exception needs to be mentioned, however, and that is the foreign-born and reared persons whose environment, though stimulating enough, was entirely different in most respects from the cultural setting in which the vocabulary test was developed. Given superior intelligence, this person can obviously acquire English vocabulary readily over a period of time, but the length of his contact with American culture is likely to affect his score. If the test were in his own language and if it involved concepts that were a part of his cultural setting, his rating would probably be truly representative of his native intelligence.

A test of so-called "verbal intelligence" should be broader in scope than mere vocabulary items, but even with several other kinds of material included it cannot be said to be culture-free. Paper-and-pencil group intelligence tests are particularly likely to emphasize the verbal or language abilities. For academic prediction, this is as it should be, but there are many jobs that call for performance rather than verbal abilities. For such jobs, the inclusion of a vocabulary subtest would be inappropriate unless it related only to terms used in the special line of work in question.

First, the worker in test development must decide whether a vocabulary exercise is appropriate for his purpose in any form. In working with employment situations, the job specifications should be helpful here along with careful observations by the examiner. In an academic setting, the teaching objectives will dictate the function of vocabulary. Then, if it is appropriate to include, the examiner must determine which of two types of exercise would be valid. General lists in intelligence tests are drawn from such fields as science, literature, law, art, music, and philosophy. They generally begin

with easy, everyday words and become progressively more and more difficult. On the individual tests, the practice is to begin with a few very easy words to "warm up" and build self-confidence. The test is then continued until five or six consecutive words have been failed. Group tests are in multiple-choice form, and they also begin with easy words and end with difficult ones, but every examinee takes all of them.

The multiple-choice form has the advantage of speed of scoring, which is important when a large volume of work is to be done. It is also completely objective, which, as mentioned before, is not entirely true on the individual forms on which the expression of the idea by the subject must be evaluated. On the other hand, of course, it is often of interest and of value to know how the individual expresses himself. The objective form has limitations here. The quantity of grading called for by the situation will determine the examiner's choice of manner of presentation.

Matching words and definitions would be another acceptable form for the presentation of vocabulary. Examples were treated in Chapter 3, and no further illustration need be given here. Where matching or multiple-choice seem equally suitable, the selection of plausible wrong answers will constitute the main problem in either case. Some examples will make this clear.

INSTRUCTIONS: After the number of each question you will find a word followed by four more words lettered A, B, C, and D. Among these four you are to find the *one* that means most nearly the same as the first word. Print the *letter* of your chosen answer in the space provided for answering the question. This answer space is at the left of the number of the question. Look at the sample below:

C SAMPLE: Start A end B complete C begin D contain
Begin means the same as *start*, and the letter C is therefore printed in the space at the left. Indicate by letter the best answers to the following:

- | | | | | |
|----------|------------|------------------|------------|-----------|
| <u>A</u> | 1. Explain | A clarify | B defend | C contest |
| | | D exaggerate | | |
| <u>D</u> | 2. Fiscal | A administrative | B physical | C exact |
| | | D financial | | |

In these illustrations, the distractors show two kinds of similarity to the first word. In number 1, choice B might come somewhat

close to the correct meaning, but it does not come as near to *explain* as A does, and therefore *clarify* is the best choice. The greater departure of C and D makes these easier to avoid. The second kind of similarity is found in item 2, in which *fiscal* and *physical* are somewhat alike in sound only. This resemblance is entirely superficial, of course, but it results in failure on the item for some individuals who, lacking any basis for a correct choice, will grasp at any clue that might prove promising. The inclusion of such false clues cannot be said to be an unfair form of trickery, since anyone who thinks deeply at all will have no trouble in avoiding the superficial resemblance if he knows the real meaning of the original word. Probably the inclusion, in some tests, of distractors much more difficult than these would be justifiable, whether to some extent they were like the first word in meaning or in sound only, or both.

TO "ESSAY" OR NOT TO "ESSAY"

IT IS characteristic of the human race to be wildly enthusiastic about every new theory that appears to work. If a new discovery is found to have value, there will be many people who think that it contains all the answers to all problems and that it should entirely replace older views. The originator of the theory or discovery may or may not share this opinion. If he is a true scientist, he will recognize some limitations to his "latest" or "new" theory. His fellow scientists will then probably point out others, and so on. If he is not scientifically trained and not very objective in his approach, he may see no value in any other theory in the field except his own. Some publications regarding new drugs to cure colds have led the public, or at least a considerable segment of it, to believe that a true preventive for all such ailments is at last available. The limitations of these drugs, valuable as they may be, have been gradually accepted as disillusionment has taken place.

Critique of Tests. The objective test has been sold as a new invention that answers all the objections to the traditional, or essay, test. It has been represented as the "final" answer to tests of all kinds—clinical, educational, diagnostic, and others. Those who are familiar with some of the advantages of objective items have tended to abandon the traditional forms altogether. Few teachers or personnel technicians have thought through the problem far enough to discover whether there are ways to make the essay question reach a reasonable standard of objectivity. Few have improved it to meet other objections that have been made. This chapter will be devoted to the consideration of the thesis that essay questions, when planned carefully, still have a useful function.

There are still many teachers who use the traditional form of examination almost exclusively. Some of them justify this practice

on the grounds that the subject matter they teach does not lend itself readily to the newer type of test, and others argue that the attainment of their teaching objectives is not properly evaluated by the use of so-called "objective" tests. A few mention time economy as the reason for using the new "objective" types. That their limitations in the skills required to prepare suitable objective items have been in part responsible for this conclusion is in many cases obvious. Those who cannot go beyond simple recognition of factual material in a true-false or multiple-choice form would fail to see that anything but essay tests could really provoke thought or require the higher mental processes for a satisfactory answer. On the other hand, objective items have their limitations. Although the foregoing discussion would seem to indicate that their scope can be quite broad, authorities in the field of mental measurement recognize today that there are aspects of learning and thinking that these tools do not measure adequately. Qualitative evaluations are being recognized more and more today as an indispensable supplement to the quantitative or objective values in tests of intelligence and personality. The way in which an individual verbalizes his ideas has been found to be of diagnostic clinical significance in the Wechsler-Bellevue individual intelligence scale and in the projective techniques for the study of personality. Critics of the qualitative analysis of test findings are eliminating from the picture of the individual tested much valuable information. That a reasonable degree of agreement can be reached on the qualitative interpretation of test responses by well-trained and experienced examiners has been demonstrated through the recent extensive use of the Rorschach ink blots and the Thematic Apperception Test pictures. The amount of subjectivity that still remains in these methods, especially in the latter, leaves much to be desired in the way of further standardization. Yet these tests are gaining acceptance by more and more psychologists with extensive training in scientific approach. While projective techniques are becoming increasingly objective (quantitative) in some ways, the qualitative is not being abandoned but is being enriched and made more useful.

Advantages of Essay Questions. Perhaps the most important reason for retaining essay items, in appropriate setting or context, in school or personnel examinations, is that no substitute has been found that will evaluate the qualitative aspects of verbal expression

of thought. To overlook the way in which a person communicates, or attempts to communicate, ideas would result in very incomplete evaluation. The consequent selection of personnel would permit many inadequate employees to get into a wide variety of responsible positions. The subsequent failure of those persons would be a critical reflection upon those responsible for the selection.

Imagine a probation officer who knew many facts about legal matters but who could not explain the rules of probation in a clear, concise manner that could be understood by the offender against the law. How would an administrator who could not write an acceptable business letter ever make policies clear to anyone else outside his own office, no matter how well he understood them? Literary style may be difficult to define and more difficult to measure, but logical presentation of ideas and an interesting way of speaking and writing may be great assets, and even absolutely essential in some situations.

Although the rare forms of objective items discussed earlier could evaluate knowledge of the mechanics of English such as punctuation and word usage, and although multiple-choice items could be made to measure critical judgment of some aspects of so-called "style," nothing can measure the creative aspect except a chance to create. No objective device for measurement of originality has yet been invented. Originality combines many things—aptitudes, stimulus, and opportunity. There is much room still for development of reasonably objective scoring criteria for the evaluation of the finished product in various fields of creative work.

From the student's or job seeker's point of view, there is often much pleasure to be derived from taking an examination that allows him the latitude to develop his thinking in his own way. This is not true of every person, but those with facility in creative writing and original thinking will welcome the chance to display whatever talents they have (or believe they have). For some kinds of employment, however, skill in self-expression would be beside the point, and in some academic courses it should be only a secondary consideration. In such instances objective questions may be all that is needed.

Wherever conveying ideas in an original, well-organized, and clear fashion is a major goal, even the outlining items discussed in

the last section of Chapter 3 fall short of the kind of evaluation desired. More than that, the objective items that require organization of thinking will not motivate students of science, for example, so strongly as will the essay form if they are really interested in new research. Educationally speaking, examinations are a part of the learning process. Learning must be motivated. Therefore an attempt to make tests a pleasure and an opportunity rather than a trial and an ordeal are worth while.

Another advantage of essay examinations is that they encourage better study habits than objective tests generally do. The former ordinarily demand recall rather than mere recognition, the learning of facts in their relationships and applications rather than in isolation, and thinking on a higher level rather than rote learning. Objective tests at their best may require some application of knowledge to practical problems, but probably not to the extent that the best forms of essay questions do. The quality of thinking that may be demanded by the grader of essay questions can be very high. Subjective factors that affect grading are somewhat offset by the thorough evaluation of thinking that is possible on essay answers but not always possible on objective answers. The student who has learned something about the grader's approach in evaluating essay answers will plan his study accordingly. If he is capable, he will accept the challenge, show his initiative, and work hard.

Most teachers recognize that the kind of question which pleases one student will displease another. They would probably agree that the most thorough study is encouraged by a variety of types of items, both objective and essay. Therefore they would usually use both on the same examination, varying the form from time to time.

An advantage that is more apparent than real is ease of construction of essay tests. It shortens considerably the time taken to write the test. Poorly planned questions may take longer to grade, or they may even have been prepared in such haste that scoring the answers is almost impossible by any system that would be considered fair. The writer was told some time ago that a history professor forgot to prepare his final examination until the hour for it had arrived. He then walked confidently into the room and wrote on the blackboard:

1. Outline the chapters.
2. Fill in the details.

This was the final examination. Less extreme illustrations are numerous. A few simple questions can cover a broad field and keep students busy for a long time. Actually the amount of time required to think through a good series of 10 specific, readily graded essay questions is usually considerable, though not so great as would be required to cover the same subject matter in 50 objective questions.

One advantage of essay questions over the objective type that is seldom mentioned is the increased difficulty of cheating when a large amount of writing is required for an answer. In some testing situations, this fact would make no difference, but in others it would reduce the motivation to borrow information. A number or letter can be copied in a flash while the instructor's back is turned. No matter how many monitors watch a large group, some cheating may, and often does, go on. Groups of students sometimes cheat for no other reason than to show how they can "outsmart" the faculty by developing a system of secret signals of such subtle nature that an expert detective would have a difficult time locating the offenders with certainty. In some schools where the honor system is supposed to prevail, it may be truly said that the professors have the honor and the students have the system. There are many facets to the problem of dishonesty on examinations, but one remedy that falls within the scope of this volume will be suggested here. Essay questions often discourage and sometimes catch the cheater. Copying an entire paragraph from another is possible under certain conditions, but it is more likely to be detected than borrowing a number here and a letter there with an occasional brief glance. The use of essay questions exclusively as a safeguard against cheating would probably not be the best solution to the problem. Such behavior is probably most effectively controlled by group pressure. Disapproval of fellow students has sometimes been known to clear up a situation in which there were a few persistent offenders. Conditions of test administration should be favorable to maintaining honesty. Alternate forms of a test often help.¹

Disadvantages of Essay Questions. As ordinarily used, essay

¹ One simple scheme which practically stops cheating is to have the students write essay answers for the first half of the examining period. Then redistribute the papers and have the students grade their fellow students. In scoring, deduct (1) their errors in their own papers and (2) the errors not corrected or incorrectly marked in the papers corrected by them. In any case, the instructor must read all papers.

items have low validity. This might not be true if a careful key for grading were made and rather rigidly followed, but usually this limitation operates without the grader being aware of most of the factors that enter into it. Frequently the grader has given little consideration to the weight that he is going to allow to factors other than content, such as logical organization, good word choice, inclusion of details, etc. Therefore he is affected by these and many others to an unknown degree.

There is no statistical formula to correct for bluffing. If an examinee does not know the facts required for an answer, he may attempt to fill considerable space with words which, when examined superficially, might appear impressive but, when studied carefully, reveal no insight into the problem. A student who admits that he does not know is probably superior academically, but he is likely in so doing to lose credit which the bluffer will receive, at least from some graders, without really deserving it. There is probably a very low positive correlation between length of answers and their quality. Often the ability to summarize briefly is an asset, and perhaps, in appropriate setting, it should be considered so by graders.

An irrelevant factor that sometimes lowers validity is handwriting. A grader who has to labor twice as long as usual to decipher the mere words of a sentence will be unlikely to grasp the central idea, even though one is definitely there. The same sentence written legibly might earn twice the rating of the former simply because the grader's attention could be focused upon meaning rather than upon mechanics of reading. If neatness as such is deliberately excluded from consideration in grading, the legibility factor still operates to an unknown degree on some papers that fall below a minimum standard. Limiting the time for answering essay material does not help to eliminate irrelevant factors. Hurrying may decrease legibility, and the fast writer will have an unfair advantage over the person who is habitually slow to the extent that, with his best efforts, he cannot cover more than half as much as the average student in the time allowed. Slow writers may not always be slow thinkers.

Spelling may be important in some courses and for some jobs, but giving it too much weight in grading physics papers may lower validity. A student who knows physics and can think clearly in that field, but who cannot spell, will argue, perhaps with some justifica-

tion, that he deserves a good grade in physics. To count spelling too much against him will make the test measure something entirely out of the field of physics achievement. Whether any phase of the mechanics of English should be counted in grading papers outside of English courses is still a debatable question. In so far as those who do not have a good command of English are examinees, essay questions are decidedly adverse. A Ph.D. from Poland might be at a very great disadvantage both in understanding the questions and in expressing his answers in English, although he would score very high if the examination were in Polish. To a lesser degree, the examinee with limited formal training in English would be affected.

Validity of essay questions is also affected by the small sampling of material that can be covered in a limited time. There is much more adequate coverage in an hour of answering time if the subject matter is converted into objective rather than essay questions. Chance plays too great a part in determining whether five essay items will hit the main points stressed in study or some minor points missed or imperfectly learned. Fifty objective items will constitute a more nearly fair sample of a broad area of material, as a rule. Of course these rough estimates of the number of questions that can be answered in an hour are subject to wide individual differences and considerable variation, depending upon the structure of the items of each type. However, in general, it is true that, for thorough coverage, the objective test saves time when compared with the essay test.

A second group of factors, overlapping the first group to some extent, tends to lower the reliability of essay tests. The most detrimental of these is subjectivity of scoring, and fortunately there are ways of improving grading that will offset this disadvantage to some extent. These devices for approaching reasonable objectivity will be discussed in detail later. A number of studies have been made of the same papers graded by several teachers or other subject-matter experts. These studies cover a wide variety of fields. All show wide variation in the ratings, even when a scale that is more or less uniform has previously been defined for graders. Sometimes one grader is consistently on the lenient side while another tends uniformly to be more exacting than the average, but more often there is a notable lack of any uniformity of pattern. Defending the extremely high or extremely low scores seems to be difficult, and the grader is apt

to resort to rather vague impressions rather than specific descriptions of his method of grading.

Many of the factors that contribute to the subjectivity of rating are unconscious. The length of an answer is one that is often found to operate below the conscious level. If the grader's attention is called to this trend in his work, he is utterly surprised, and indeed often reluctant to admit that he could be using length as a criterion, even if the evidence stacks up against him. Omit from consideration, at least for the present, the grader who deliberately rewards a verbose treatment of a problem. There still remains the well-meaning teacher who unknowingly allows himself to be carried away by padding, or unnecessary words included in the answer. It is a difficult fault to correct. The ability to summarize briefly is a virtue and a skill often needed. The grader should recognize it and know how to reward it.

The general feeling of well-being or the opposite mood in the grader is a normal human phenomenon that is difficult to deal with in essay evaluation. This variable is much discussed by high school and college students, who are quick to recognize in a teacher moodiness that borders upon the extreme or abnormal. Some teachers frankly admit that some of their errors are the effect of mood and try to correct them, but others are unaware of any such influence upon their rating. The halo effect is closely related to this problem. When a teacher is in a pleasant mood while grading a paper, he is likely to find a good answer to question 1 and carry this impression over to question 2 on the same paper, and so on, thus unconsciously giving the student too much credit for an inferior answer now and then. It is equally possible to be the victim of what some authors have called the "horns effect," that amounts to carrying an impression of an inferior answer over to all the rest of the work a student has done, some of which may be fairly good in quality. Although the horns effect is not nearly so common as the halo effect, either one can occur without the grader's being aware of its subtle influence upon test scores. Doubtless the cumulative effect can become considerable.

Avoidance of the effect of the mood of the grader may not be entirely possible except in dealing with answers of a factual nature for which a detailed scoring key is prepared in advance. Its influence can be reduced if a detailed key is followed whenever possible. The

halo effect and horns effect are best handled by grading all answers to question 1 first, then proceeding to question 2, and so on. Finally, the grader must look at the examinee's name, of course, but when possible he should avoid doing so during the process of grading. In some situations the identity of the examinee would mean little if anything to the grader, but in others the past performance on other tests would have some influence, perhaps at the unconscious level, possibly clearly at a conscious level. Familiar handwriting may occasionally reveal the identity of a student well known by his professor even if the latter does not look at the name, and for this there is no known remedy except typing of answers by all students. While typing would eliminate the marked differences in legibility and neatness of handwriting mentioned earlier, this would not always be easy to arrange in an essay-test situation.

Less important factors that may sometimes influence grading to a marked degree include fatigue of the grader, contrast between successive papers that tends to exaggerate the differences between them, and the value that the grader places upon accurate and honest scoring. Many teachers place little importance upon tests, and merely go through the motions of giving them in order to satisfy those in authority who may expect them to give quizzes and examinations. Still others regard these academic devices as mere exercises to give the student practice with material he is learning rather than primarily as means of evaluating progress fairly and exactly. Perhaps it is mainly attitudes like these which have resulted in the extreme unreliability of grading reported in some studies.

While space does not permit detailed review of the numerous experimental studies on reliability of essay questions, mention should be made of the fact that many of these represent essay tests at their worst rather than in improved form. As ordinarily used, essay tests may have low validity and low reliability, as evidence from most of these studies has clearly shown, but there are improvements in methods of grading that may significantly decrease the influence of some of the disadvantages listed here. This point will be developed further in the last section of this chapter. It might also be said that research has been largely concerned with the reliability of the grading of essay tests rather than with the reliability of the tests themselves. The construction phase and the grading phase are both im-

portant and distinct skills in which some individuals have a greatly exaggerated notion of their ability and achievement. Neither of these skills is as simple as it may appear on the surface. Perhaps workers with tests can be motivated to take each of these tasks more seriously.

Test Forms and Examples. In some ways, the essay question may overlap the objective form which in this volume has been termed *completion*. From the simple recall type, the essay item progresses in difficulty by stages to include the most challenging problems, requiring original and creative thinking in its most carefully planned form. Several attempts have been made to name these stages, and the summary here is an attempt to stress important distinctions rather than to cover every conceivable formulation that has been suggested.

Students at high school and college level have been observed by the writer to be far too often careless or perhaps misinformed about the meaning of important terms found in questions from every subject-matter field. Even superior students frequently *discuss* or *describe* when asked to *define*. They may also *elaborate* in great detail when asked to *summarize* or perhaps to *list*. They may *list* when directed to *define* or *distinguish*. It may therefore be advisable, if not imperative, to underline or even to explain these terms when they are used on an examination. To show the broad scope of problems that may be presented and to distinguish various basic forms, the following summary is presented in approximate order of difficulty.

1. *List*. In its simplest form, this item is rather objective, similar to completion, in that it requires recall of factual material. It does not call for evaluation, relationship between facts, or organization in any particular arrangement. It can usually be graded with a simple key with a high degree of reliability. Some might question whether it fits the definition of an *essay* item.

EXAMPLE: List three countries which at present use the Soviet system of Communist government.

2. *Arrange*. Not only is recall demanded here, but also a given order of recalled facts that is specified. The basis for the order (chronological, increasing importance, decreasing value, etc.) is an essential part of the question. The mental process is not quite so simple in most of these as in type 1 described above, and often a rather difficult basis for the arrangement is given.

EXAMPLE: List, in order from least reliable to most reliable, three measures of variability.

An acceptable answer would be *range*, *quartile deviation*, and *standard deviation*. Recognition of this as a correct sequence requires more knowledge than just the names of three measures of dispersion.

3. *Select*. On a given basis, list items that belong in a category. This is usually simple evaluation that amounts to picking out the most important, the least valuable, the most controversial, the most recent, etc. This type may be more difficult than type 2, or it may be easier, depending upon the nature of the subject matter.

EXAMPLE: Which of the presidents of the United States served while this country was at war?

This is an item depending upon factual knowledge. It involves sets of facts that have to be related in a clearly defined way and classified in order to answer the question. Another formulation that belongs under this same heading would require evaluation of material rather than mere classification.

EXAMPLE: Name three scientists who have made outstanding contributions to atomic theory.

Although some answers might create controversy among several expert graders, others would be accepted by all under the loosely defined concept of an "outstanding contribution." Both answering and grading would in this case present more difficulties.

4. *Describe*. The task here is to give the important characteristics of an object, process, or phenomenon. The item may call for either a brief or a detailed description, but in either case the answer goes beyond a mere definition. A paragraph rather than a single sentence is usually appropriate. In some ways this may be easier than giving a concise definition, as will be shown later.

EXAMPLE: Describe an amoeba.

The wording may just as well be in the interrogative in such a form as, "What are the characteristics of the amoeba?" In either case the animal's structure, mode of locomotion, digestion, method of reproduction, etc., would be expected in the answer.

5. *Discuss.* The examinee is required to go beyond description into full elaboration of the subject with tracing of development, arguments for and against, and perhaps relationships to other facts or ideas, depending upon what has been learned. This form of item will vary greatly in difficulty depending upon the context in which it is used. Its position in this series may therefore be a matter of debate. Certainly there is no form that gives more latitude as to approach. Although the variety of ways in which an answer can be treated may sometimes lead to rather frustrating problems in scoring, the individual differences in originality and in organizing ability that are revealed will often distinguish clearly between a superior person and a mediocre one. The frame of reference is very important in answering questions of this sort. Wording will have much to do with comprehension of the approach desired. Unless the question is in itself very specific, or unless the students know in advance the particular slant favored by their instructor and/or grader, the variability in material presented as an answer and in the treatment of the material will be enormous. If a question in this form can be made to require a very specific treatment without giving away the answer or curbing originality too much, the scoring task will be made much less of a burden.

EXAMPLE: Discuss the rehabilitation of juvenile delinquents in an institutional environment.

In this example the area is somewhat restricted. It does not include all offenders against the law, nor does it include all juvenile offenders, since many are not institutionalized. Yet it leaves room for the well-informed student to develop a topic that is full of possibilities. Originality and clear thinking can be displayed in its development.

6. *Define.* This means essentially to place in a category and distinguish from everything else. This is often more difficult than discussion, though not always. Many students when asked to define are inclined to describe or discuss. Another way of explaining definition is that it states the limits of a subject. This should be done briefly, but clearly, not vaguely. Skill in defining terms is essential not only to research but to an understanding of reading matter and

to communication of ideas in general. An illustration may be taken from the field of statistics.

EXAMPLE: Define *median*.

An acceptable answer, though not necessarily the very best one, might be, "The median is a measure of central tendency or average on either side of which 50 per cent of the cases in a distribution fall." This answer first places the term in a general class, as a kind of measure of central tendency or average.² Then it distinguishes the median from the arithmetic mean or the mode by stating how it is located in the distribution. Probably most, though perhaps not all, definitions tend to follow this pattern. A vague statement that omits either of these two steps (categorizing and distinguishing) probably falls short, in most instances, of the standard of acceptability.

7. *Illustrate*. To ensure that rote memory is not being used on definitions, laws, and principles, it is useful to call for an example. This is the real test of the learner's ability to apply his knowledge.

EXAMPLE: Give an illustration of the Doppler effect.

This phenomenon in physics of sound is often experienced when a train on which the observer is riding approaches a crossing bell. The pitch of the bell grows higher as long as the observer is approaching the source of the sound and becomes lower as he moves away from the source of the sound.

8. *Explain*. An explanation places special emphasis upon cause-and-effect relationships. The difficulty of the item depends almost entirely upon the context. Wording may vary considerably.

EXAMPLE: Why are relatively fewer victims of "polio" permanently crippled than was the case a generation ago?

The presentation of several important facts about diagnosis, treatment, and bodily repair would be essential here. Often the answer to this type runs into several pages of material.

9. *Compare*. This form may or may not state a specific basis for comparison, but in either case it is likely to call for the advantages

² Another definition might be, "The median is the middle item of an array of things," or, "It is the position of the middle item in an array." Its "size" may cause it to be an indication of central tendency.

and disadvantages of two ideas, their similarities and differences, etc. A good answer should show a well-planned approach rather than a haphazard one. The broader the question, the more difficult will be the scoring problem. An example of a general question is given below.

EXAMPLE: Compare the objective examination with the essay type of test.

A rather lengthy answer would be required to treat this fully, but it might be shortened considerably and made easier to score if it were limited in some way, such as:

EXAMPLE: Compare the objective examination with the essay type of test from the standpoint of scoring.

10. *Summarize*. This means most nearly to state in as brief form as possible the essentials of a unit of material. The term *review* may come close to the same meaning, especially if the question begins, "Briefly review."

EXAMPLE: Summarize the events that led up to World War II.

11. *Outline*. This is somewhat related to *summarize*, but usually implies organization into main topics and subheadings, whereas summarizing may be done in a short paragraph.

EXAMPLE: Outline the development of progressive education in the United States.

Tracing this movement through several steps or stages, naming outstanding thinkers in the field and their contributions, would probably be satisfactory, provided that not too much detail were given. *Discuss* would differ from *outline* chiefly in the amount of elaboration involved in discussion, and perhaps in the inclusion of arguments for and against, which would be out of place in outlining.

12. *Interpret*. An extension of *define* is found when the meaning of a rather profound or obscure quotation is desired. Reading comprehension and word fluency both enter into successful performance of this task. This form of item may also call for practical application of general principles. Mere rote memory will not be enough for this task.

EXAMPLE: How can the gestalt school of thought be applied to the reading of a printed page?

To answer this, one not only must be familiar with the literature on the gestalt school in psychology in a general way, but must be able to show how a word or phrase is organized into a perceptual whole that is immediately grasped.

13. *Criticize.* To evaluate the correctness or adequacy of an idea, usually also giving suggestions for improving it or reasons why it should be abandoned, is to criticize. Popular opinion is to the effect that it is easy to criticize almost anything, but to have any value, criticism should be well thought through. Broad knowledge and even experience may often be required for evaluation to be worth anything.

EXAMPLE: Criticize the practice of learning a foreign language primarily by translating it into English.

14. *Formulate new problems or procedures.* The originality and ability to think creatively with the tools and materials of a given subject-matter field make this perhaps generally the most difficult of all types of essay items. Grading is necessarily largely subjective, yet it is the outstanding answers to these items that often reveal aptitude for research that merits further development.

Approaching Reasonable Objectivity in Scoring. In a teaching situation it is always good practice to help students learn to use the most efficient methods in preparing for examinations. This does not mean that the teacher should give the answer to every question that later will be asked. In fact, such a procedure would not encourage thorough study of the right kind. The statement really means instruction in study habits and in the interpretation of questions of various types, such as those discussed in this volume. In dealing with the essay forms just discussed, the teacher may increase the reliability and objectivity of grading by impressing upon students important distinctions between *define* and *describe* and other similar pairs. Also he may show the advantages of well-planned answers over haphazard ones containing the same facts. He may explain how much extra credit, if any, he plans to give for logical presentation, clear expression, and other factors besides the basic points of

information required. Such explanation will probably tend to bring his students to regard tests as a legitimate part of the learning process rather than an unfair and unnecessary ordeal. If, later, his students should compete with others for jobs for which they were well qualified, and if civil service examinations or industrial tests should be a part of the employee-selection procedure, they would have one decided advantage in their preparedness to handle a test situation intelligently.

An example of lack of efficient method in approaching an essay question may frequently be seen in the student who is in a hurry to finish. He wastes rather than saves time by beginning to write immediately after reading the question, never stopping to think how he might organize his answer or what approach to it might present the matter most effectively. The result is a mass of details with no important points standing out from the rest. The sum total may represent only two-thirds of what he actually knows that relates to the item he is answering. Again, it may represent a complete misinterpretation of the question through mere carelessness. A specific instance of the contrast between such a student and one who organized his thinking well comes to mind at this point. The two were about equal in the amount of factual information they were gaining from their courses, but student A recalled each point in relative isolation from anything else, while student B depended upon logic and careful planning in order to make sure he put down on paper everything he knew that would help answer the question. The difference between their scholastic attainment was not so evident from objective tests, especially those in which rote memory played a major part. Their scores on these were nearly equal except that B did better on multiple-choice items requiring original thinking or application of principles.

It was on essay answers that B showed his marked superiority to A. In one undergraduate course in psychology, both students had attempted to answer the following: "Compare the Wechsler-Bellevue with the Revised Stanford-Binet as intelligence tests for clinical use." The answer given by A is as follows:

The Wechsler is more diagnostic than the Stanford-Binet is. The Stanford-Binet is not so good for adults. It is an age scale, and the Wechsler is a point scale. The Wechsler has material that is of more interest to most

adults than the Stanford-Binet is. They are both good individual tests. The Wechsler has more performance in it, while the Stanford covers a wider range of ages.

This rambling chain of free associations would have to be read at least twice to find out how much of the essential material it did contain. Actually the essentials are largely present so far as the material of the course developed them, but the wording is awkward, and anything remotely resembling planning is absent.

The answer given by B is reproduced below.

Both of these are individual intelligence tests, but the Stanford-Binet is an age scale, while the Wechsler-Bellevue is a point scale.

The Stanford-Binet has the following advantages:

1. It is better suited for children.
2. It covers a wider age range than the Wechsler-Bellevue.

It has the following disadvantages:

1. It has insufficient nonverbal material at higher levels.
2. Some items in it are fast becoming obsolete.
3. The adult-level material is not interesting enough to most adults.

The Wechsler-Bellevue has the following advantages:

1. It is more diagnostic than the Stanford-Binet.
2. Performance as well as verbal scores are obtained.
3. The material is interesting to most adults.

It has the disadvantage of being unsuited for younger children.

The systematic presentation of B's answer makes the task easy for the reader. The gestalt is easily grasped at the first reading. Although the arrangement of topics on the page may have some advantage, the same words without numbering and all run together into one paragraph would still be decidedly superior to A's answer. The facts mentioned are almost identical, although B managed to give a little more. Grading on ideas, one point for each idea, the writer would credit A with 8 points and B with 12. However, he would also credit B with 1 point each for (1) grouping advantages and disadvantages, (2) grouping likenesses and differences, and (3) clear wording of ideas. The total would then be 15 for B.

These and other similar pairs of answers have been given by the

writer to graduate students in a class in test construction. The instructions were to assign ratings to the two answers and justify the procedure used. Results showed for a small group of graders that over 50 per cent did not spontaneously grade on any other basis but content. Some of these later agreed, however, upon comparing notes with the others, that organization should have some reward. The amount of such credits and the scheme for arriving at them varied greatly. Experienced teachers also seem to differ in their opinions on this matter, but little actual research has been found by the writer in this area.

In order to improve and objectify the grading of essay answers, some degree of planning regarding factors other than content seems essential. A prepared key, therefore, should allow for such factors in a planned way, if they are to be considered at all. Before beginning a key, the grader should think about the purpose or purposes of the question. With definite objectives in mind, he can then work out a tentative key, specifying the number of credits for each idea or factor. The entire plan may prove inadequate when the actual papers have been inspected, and the grader must be willing to change the plan if circumstances clearly justify his doing so.

Rinsland³ makes a distinction between scoring and grading. Scoring is the objective process of counting right and wrong responses, while grading involves the qualitative interpretation of material in terms of a criterion. Both of these approaches are needed in essay evaluation if the objectives of essay questions as outlined here are to be achieved.

In order for an adequate key to be possible, the question itself must be specific. It must limit the field and the approach, to some extent. Many items with answers a paragraph or two long will yield better results than a few very broad items requiring lengthy answers. If the answer key usually turns out to be unworkable, there is probably something wrong with the wording of the questions or else with the training of the examinees in the skills of taking examinations. The former hypothesis should usually be investigated first.

Making a Key for Scoring. A good procedure in making a key might include at least the following basic steps:

³ Rinsland, Henry Daniel, *Constructing Tests and Grading in Elementary and High School Subjects*, New York: Prentice-Hall, Inc., 1938, p. 302.

1. List the main points required for a satisfactory answer.
2. Weight these points, or assign them numerical values that seem appropriate to their relative importance. These usually are in terms of per cent, but they need not necessarily be graded on that basis. A raw score can be converted to per cent later.
3. Allow room for extra points to be given if the material is unusually well elaborated or if the organization of the material shows exceptionally logical development. Some questions would be of such a nature as to make these credits for elaboration or development extremely important, while others would be so simple and factual as to justify credit for main points only. The amount of such credit allowable on any question, of course, will be a matter of opinion.
4. Test out the key for adequacy by grading all papers in a group on that particular question before going on to the next question. (In doing this, the grader may discover some points or approaches that did not occur to him originally, but which he will recognize immediately as deserving of credit. If he thinks critically at this point, he will probably not be too liberal in changing his key to allow for these new approaches or points.)

In applying a key, once it is made, there may be little difficulty in determining the number of facts that are in the answer. The qualitative aspects of presentation are likely to give the most trouble. Here the suggestion has often been given that few rather than many degrees of merit should be established. It may be possible to achieve a fair degree of reliability with 3 to 5 levels of excellence, but not with 10. One procedure would be to sort papers into 5 groups. If the entire group followed the normal distribution curve, one would expect them to approximate the following percentages: very superior, 10 per cent; superior, 20 per cent; average, 40 per cent; inferior, 20 per cent; and very inferior, 10 per cent. Not every group of papers, even though quite large, will fit this distribution, however, and inaccuracies would result from an attempt to force them to fit the normal curve. A careful rereading may be necessary to shift any papers that have been misplaced.

If a skewed distribution occurs (one which is piled up near either end rather than in the middle), the examiner should try to determine

the cause of this skewness. If he is acquainted with the elements of the theory of probability, he will doubtless recognize several possibilities. He may have a select group that is uncommonly high or exceptionally deficient in whatever trait he is measuring. If he knows the background of his group well, and if he has had frequent measures of their past performance, he may be able to explain his skewed curve. However, he may reach the conclusion that he has a normally distributed group but that this particular test is too hard or too easy, or that he has graded too leniently or too severely. As long as he has been consistent in every way possible, he may not have to reread every paper. He may simply have to reset his passing point and convert to per cent on a new scale.

Possibly an example of a key in an actual test situation would clarify some points of procedure. While the illustration to follow was intended to apply to an undergraduate course in which the students did not have the background to develop the topic fully, it covers the points presented to the students in their textbook.

QUESTION: What would probably be the effect of sterilization of mental defectives upon the percentage of the population that are mentally deficient?

ANSWER:

1. Sterilization would probably have very slight effect.	<i>Credits</i>
a. Many carriers of hereditary feeble-mindedness are themselves normal individuals.	5
b. Feeble-mindedness may result from birth injury or other causes rather than heredity in many cases.	3
c. Administration of sterilization would not be anywhere near possible with limited number of qualified psychologists to give tests.	3
d. Political abuses, moral and religious objections would interfere with fair administration even if legislation passed.	3
<i>Extra credit:</i> Clear organization of thinking.	3
<i>Total</i>	3
	20

In this key, the greatest weight goes to the main point made by the textbook that little effect would probably be shown. This would receive 5 points. The four reasons for this conclusion would then be less important, and could be given 2 or 3 points each. Three was

selected so that, with extra credit for good presentation, a total of 20 would result. Although some variety of angles of approach did occur in the first group who answered the question, the key remained in this form after tryout. Some stressed the legal angle, while others elaborated the biological.

This chapter concludes with the suggestion that further research be performed with trained graders using a key such as might be prepared using the principles described here. Work with essay questions has convinced the writer that greater reliability can be demonstrated than has been achieved in many previous studies of the essay test. In the usual case, the plan can be worked out jointly by two and sometimes three graders. After independent scoring, the graders should compare notes and reconcile differences, as far as possible. This is especially important where grades are given to examinees, a practice which is certain to result in critical comparison among themselves. A good test should not show more than 5 per cent of the answers with wide differences, of 10 per cent or more per answer, in a sample of those examined. No graders are perfect, and essay questions have too much subjectivity to be satisfactory in every respect, but they can be made to reach a reasonable standard of objectivity under good conditions.

If more than two or three graders are used, what is known as independent, or multiple, grading must be used. That is, each paper must be graded by more than one grader. If the difference between the highest and lowest grade by any grader is greater than 5 per cent, some review and reconciliation becomes necessary before a final grade is assigned to an examinee.

PERFORMANCE TESTS

MOST high school and college teachers of academic subjects are inclined to think of tests in terms of either (1) the appraisal of acquired knowledge, (2) information, or (3) aptitudes, largely of a verbal nature, the measurement of which will enable one to predict how successful a person will be in acquiring knowledge. Teachers of commercial subjects or skilled trades and specialists in personnel selection are clearly aware, however, that there is another broad area of human behavior that needs intensive study and evaluation. This area includes both skills of various kinds and the native aptitudes, great in number, that enable us to acquire these skills. Perhaps there is no sharp line of distinction between these two broad aspects of human learning capacity, but the measurement of specific skills already acquired and the prediction of success in acquiring them are two important problems quite different in nature from the problems of verbal appraisal. Although there are skills involved in our thinking processes which help us in answering questions on a written test, these might be classified roughly under *habits of thinking*. The new problem arises in the measurement of what might be called manual, or manipulative, skills. These are largely complex systems of motor habits. Some knowledge and perhaps reasoning ability of a sort may be required in their application on a specific job, but the possession of the needed factual information does not ensure the presence of well-learned habit patterns. Without a minimum standard of motor coordination achieved through practice, the factual knowledge of how to do the job may be practically useless.

Why Some Tests Are Not Written Tests. Some skills described already in this volume are adequately appraised by means of a written test. These are in the category of mental skills or habits of thinking, such as the mechanics of English usage, arithmetical computa-

tion, alphabetic filing, etc. Motor coordination is not involved in these except to the limited degree required to write numbers and words legibly. The evaluation of progress in the training of typists, stenographers, carpenters, automotive mechanics, etc., requires the measurement of motor or manipulative skills.

It is possible for a person through intensive study to pass a written test of information about plumbing without having acquired the skills necessary to perform the simplest operations of the trade in a satisfactory manner. A civil service examination for radio repairman consisting of simply worded multiple-choice items was tried out on a physics instructor and an experienced radio repairman in Baton Rouge, La. The instructor made an almost perfect score, missing only one item, while the repairman could not score 70 per cent right, and answered only 60 per cent of the items correctly. The instructor knew his theory well. He could define *ohms* and *amperes*, and his comprehension of theoretical principles was superior. However, he later attempted to repair a radio and failed completely, perhaps for lack of practical experience, or manual skill, or both. The repairman could not identify definitions of technical terms correctly, and he did not appear to recognize theoretical principles. His knowledge consisted largely of locations of parts and how to connect them. He was skillful with tools, and was able to locate and remedy promptly the trouble with the radio on which the instructor after much time and effort had completely failed. The personnel agency constructing the written test was impressed from this preliminary tryout with the necessity of some measure of the practical ability to handle tools and radio parts.

Various names have been applied to test measures of skill. In the skilled trades, the terms *practical shop test*, *work-sample test*, and *demonstration test* have been widely used. The general term *performance test* has been used to cover two quite different types of measures. Each of these measures has developed out of a different approach. Thorndike¹ has best described these approaches as the job approach and the trait approach. The former he sets forth as developing out of the duties performed on the job, giving as an example the construction of apparatus that duplicates the flying situa-

¹ Thorndike, Robert L., *Personnel Selection (Test and Measurement Techniques)*. New York: John Wiley & Sons, Inc., 1949.

tion as closely as possible to measure the perceptual-motor responses of pilots. This is the work-sample type of performance test. It has its applications in many occupations and types of training. Its aim is always to make the materials of the learning situation as similar as possible to those which would be encountered on the job. The other approach is the trait approach, concerning which Thorndike² says that "the test development is based on the general qualities of the individual rather than on the characteristics of a specific job." Examples of performance tests constructed for such general measurement of aptitude would include the Kohs Block Design Test, the Minnesota Rate of Manipulation Test, and many others. The trait approach has the decided advantage that each of the fundamental capacities measured will probably have a very low if any correlation with every other trait. Such measurements are concerned with very basic, elementary functions, while work samples are complex patterns of human behavior. The work samples, however, have greater face validity. The carpenter wants to be tried out with the tools to which he is accustomed. He will not be highly motivated by a test of theoretical type, which, though actually it may be valid, does not seem to him to be remotely connected with his work. The problem of selecting or constructing an appropriate performance test boils down to the purpose for which it is to be used and the population who will be taking it. Paper-and-pencil tests are too far away from the job in many cases. Thorndike has found that they have been quite inadequate for the measurement of skills, and that apparatus is almost indispensable. Although tapping and aiming as well as various kinds of maze-tracing tasks have been used as items in a group situation with pencil and paper, these inexpensive screening devices will probably never take the place of apparatus tests.

Motion-picture tests may be presented in group situations if perceptual functions are to be evaluated. Perception of direction and speed of movement may be measured under quite well-standardized conditions if movies are used. The measurement of speed of reaction to perceived stimuli is a more complex problem, however, that calls for individual rating of performance. Large-scale personnel selection where skills are involved is expensive. Printed tests will do only part of the work. Not all tests can be written tests. A sample of job per-

² *Ibid.*

formance is often required. In the achievement of educational objectives in science, laboratory performance must often be evaluated along with knowledge of theory. A student who excels in theory may not always be superior in laboratory work. The reverse is equally true, because two more or less independent sets of native capacities are involved. In the teaching of trades and commercial subjects, the written test may assume an even more subordinate role. The reader is referred to Thorndike's *Personnel Selection* for a detailed discussion of the choice of an appropriate test battery for personnel selection.

Standardization of Material and Procedure. The selection of suitable material for performance tests will be governed by a number of factors. The most important one is probably whether the test is designed for prediction of skills not yet acquired or to measure skills already acquired. Measures of aptitude of this nonverbal sort are not independent of skills already acquired. They are perhaps only less specialized in content than measures of achievement in specific trades or other manual occupations. A driving test for taxi or truck drivers does not attempt to predict how rapidly a man will learn the operation of the vehicle, but rather it attempts the evaluation of his present driving skill. Only a machinist could very well be given some tasks to perform on a lathe, since an untrained person would not be able to produce a measurable performance in a reasonable length of time. A work-sample test, therefore, will ordinarily call for the actual tools of the job.

According to Adkins *et al.*,³ a work sample suitable for a performance test is seldom a piece of work taken in its entirety from the actual job. Some parts of almost any job will yield little variation in performance, and such parts, if included on a test, will not measure anything. Tasks must be selected that will differentiate between good and poor workers and that will, if possible, yield several different degrees of proficiency. Such tasks must also constitute a fair sample of the various kinds of work required on the job. For instance, several little repair operations on motors would be insufficient as a performance test for general electrician, because it is conceivable that

³ Adkins, Dorothy C., Primoff, Ernest S., McAdoo, Harold L., *et al.*, *Construction and Analysis of Achievement Tests*. Washington: Government Printing Office, 1947.

a man might know motors quite well and yet be entirely inexperienced in some of the problems of house wiring. To find tasks that meet this requirement and that will not consume too much time is not easy. An objective type of written test can cover a wide sample of duties much more quickly than can a performance test. Therefore, a compromise between high reliability and reasonable cost of administration may be necessary. Reliability must not be sacrificed too much to cut costs if anything approaching scientific measurement is desired.

Even if a sufficient number of tasks that are a representative sample of the job are selected, the examiner cannot be sure of reliability of measurement unless he can show that different scorers will agree reasonably well in scoring. Objectivity in this regard is indeed difficult to achieve, even when the scorers are recognized as accomplished in their specialties. Quantity of work may be an objective criterion that is accurately measurable, but the evaluation of quality is a different matter. The complexity of factors that enter into the final product often makes for radical disagreement among experts. Adkins maintains that even ratings of carpenters on a simple task like sawing will be done on the basis of several completely superficial and unimportant factors, such as rubbing the wood with the fingers in a familiar manner or placing a hammer in the side pocket of overalls as if well acquainted with tools. There is no doubt that conventional ways of handling tools, which have no effect upon the work, may enter into rating in the trades.

Even when the rating factors have been rather clearly defined for the examiners, much difference of opinion will often occur. Probably accuracy of the final product is often the best criterion. In carpentry, this could be achieved by giving specific instructions to examinees, for example, to follow certain specifications, using specified tools only. Accuracy of cutting then could be measured by the raters at specified points on the finished product. If the raw material were the same for everyone, and if the test were timed, there would be little room for disagreement in scoring. Not all qualitative factors are as easily measurable as the example described above, but this degree of objectivity should be the goal of test construction whenever possible.

Only a tryout will demonstrate whether a test procedure will yield

reliable results. Tasks that appear suitable for rating both to the expert in the occupation and to the test expert may turn out badly after a trial run with several subjects and several raters. When only one rater is used, as in most educational situations, even more precaution is necessary to ensure freedom from personal bias; few instructors will attain this end. Not only is personal bias apt to enter into making results unreliable, but the impossibility of control of all testing conditions may lead to failure of a proposed procedure. Unless the condition of tools and materials can be kept reasonably constant, each candidate will be presented with a different situation. Difficulty cannot be equated unless this requirement can be met. Although the sharpness of cutting tools can perhaps be regulated to some degree, the quality of materials to be cut, particularly wood, may present a more difficult problem in testing by means of a work sample upon woodwork.

A performance test may have the appearance of practicality (face validity) and still not be valid. If the tasks selected for the test are ones seldom encountered on the job itself, skill in doing tasks required every day may not be evaluated. A high score may then mean successful performance of the less important duties, but no assurance that the most essential requirements of the work can be met. Although brief operations save testing time, these may not be the critical phases of the work which enable the test to distinguish between efficient and inefficient employees. The selection of important aspects of the work to measure must be based upon job analysis, just as in the selection of written test items. Those operations which are observed to depend too much upon chance factors for their success present too much variability to make a good work sample. The examiner needs material that will tend to produce uniform results each time the person takes the test. If too much chance enters into it, the results are neither reliable nor valid.

For example, suppose that a Dictaphone operator examination was given, using several recordings by different dictators, some of whom spoke clearly and distinctly while others did not. While the operator might ultimately be in an office where he would have to transcribe from records that were rather muffled or blurred, this situation is one to which he would gradually become accustomed. To allow variability to occur in quality of dictation would introduce

an unnecessary chance element into the work sample. Those who got clear records to transcribe would have a decided advantage. Therefore the examiner should make sure that the records meet a reasonable standard of intelligibility, and perhaps this is best achieved by having one dictator do all of them.

Whenever the condition of materials or equipment introduces an element of uncertainty that is beyond the control of the performer, the results may mean accidental success or failure rather than competence or incompetence. While such factors of chance can hardly be entirely eliminated, they can often be reduced to an almost negligible importance by careful selection of the task and thorough inspection of equipment and material. In a typing exercise, the conditions can usually be held fairly constant if good machines are available. A warming-up period before actual testing often catches defective machines in time to prevent accidental failure, if the examiner is a careful observer and instructs his examinees to report defects during this preliminary period. In other situations, the control of conditions is more of a problem. The writer once had charge of the administration of the Hand-tool Dexterity Test. Bolts and nuts were to be removed from one side of a frame with wrenches and placed and tightened on the other side of the frame. The instructions were clear and the conditions were constant except for one point—the tightness of the nuts to prepare the test. If more than one examiner administers the test to a group of competitors, it is almost impossible to have this tightness constant for all subjects. How much this kind of variable affects results in this or other situations is difficult to determine.

Study of the examples in Appendixes B and C will clarify a few of the points discussed in the preceding paragraphs. The first of these is a work sample from a battery of tests for counselors. Included in the battery were (1) a written test covering factual information about occupations, tests, interviewing, etc., mostly in multiple-choice form, and (2) a practical demonstration of counseling skill. The battery was suitable for an academic situation as a final examination in an advanced counseling course at university level. Quasi-executive and management-employee selection might make use of similar performance tests. In different forms the work-sample procedure de-

scribed here had already been tried in both the academic and the personnel setup.

More specifically, the procedure in Appendix B may be described as one in which the examiner prepared in advance a script consisting of the actual comments of a client in an interview situation presenting a variety of counseling problems. The examinee was instructed to the effect that a nondirective approach was desired. A few identifying facts were first given about the client and the reason for referral. The examiner then took the examinee into an interviewing room and read the client's part; the examinee responded to this. The student counselor's remarks were recorded on a tape recorder in order that they could be transcribed for study later. Ideally, the skill with which the counselor-examinee used his voice could be evaluated from the tape as well as the written comments. But, when the volume of such testing is large, too much time would be required for much attention to be given to this particular concept. The interview situation was made as natural as possible. Counselors' remarks were spoken rather than written, because the ability to think quickly what to say and when to say it is regarded as extremely important. In Appendix B will be found actual material used in one such work-sample test in an academic setting.

In Appendix C are complete interviews thus recorded. They are for three typical student counselors as graded finally by professional examiners. One interview is that of the examinee graded highest by the three professional examiners, one of the lowest, and one of the "average." The grading was not identical by the three professional examiners. That they had differences of opinion upon the relative merits of the examinees is of little significance beyond indicating the fact that grading of performance tests is not an exact process.

As to presentation, this material was well standardized. The test has the decided advantage of being practical, in other words, of having face validity. In so far as the examiner can make his oral delivery of the written matter constant for every individual, the situation is fairly well controlled. However, there is one uncontrolled factor in the sequence of exchanges of remarks. In a normal counseling situation, the client is affected to some extent by some of the counselor's remarks. What the client will say next is often determined

by the counselor's response to what the client has just finished saying. Therefore, since each counselor's approach may vary each time, the conversation will seem a bit disjointed, illogical, and unnatural in some parts of the interview. This variable could hardly be avoided without presenting each examinee with a different problem in counseling, and this would be entirely undesirable, of course.

The scoring of the material in Appendix B is a separate problem to be considered in the last section of this chapter, but at this point may be mentioned the fact that there are many kinds of counseling jobs presenting different problems. The material given here covers only a small sample of the problems that would be encountered in these jobs. Taken collectively there is probably at least enough for a fair indication of counseling skill. Validity studies have not been done on this material, but they should be carried out before it can be said conclusively that this is a well-standardized and useful procedure. Philosophies of counseling differ, and in order for results to be evaluated, they must be rated on the basis of their conformity or lack of conformity to some accepted school of thought, such as that known as nondirective. Otherwise, much confusion might arise over rating. Even within this frame of reference, objectivity is difficult enough to achieve.

The second illustration is the design of a performance section of a battery intended for personnel selection for positions such as psychometrician or psychological assistant. As in the previous example, a written test of information would be used with the work sample. This practical demonstration has been tried out in an academic setting as a part of the final examination in a practicum in clinical testing, but not to the writer's knowledge in actual personnel selection, at least not in exactly this form.

The materials consisted of a stop watch and materials, manuals, and record booklets for the Wechsler-Bellevue Form 1 and the Revised Stanford-Binet, Forms L and M. Other tests could just as well be included if the course or job required them. The items that were to be administered by each examinee included those that involved complex instructions with illustrations or difficult scoring problems, such as Copying a Bead Chain from Memory, Discrimination of Forms, Maze Tracing, Repeating Thought of Passage on Value of

Life, etc. (Stanford-Binet), and Similarities, Block Design, Object Assembly, etc. (Wechsler).

Each individual administered a total of 10 subtests to the examiner individually. The writer served as subject, giving responses that would challenge the best skills of the student in giving instructions and in scoring. Each student scored and explained his scoring as he went along. The responses were selected from actual case records in the writer's experience, and were performed in as nearly uniform fashion as possible. These examples of test performance were selected to represent problems often encountered in testing experience. Errors in administration and scoring were weighted beforehand as far as practicable in terms of their importance, and then, as the test was being conducted, the writer checked errors. At times he had to make note of a mistake or defect not previously encountered, and an element of subjectivity often entered in here. On the whole, however, the method was fairly well standardized in advance of administration.

In academic courses over a period of years, the writer has found by rank correlation that scores on an objective written test covering the theory and the test materials correlated only about .65 with this work sample. Students are often confused in the situation, and omit important steps, fail to think through the scoring properly, etc., even when their theoretical knowledge is quite good. Personality factors in the student examiners are likely to enter into this kind of failure.

In this section, several work-sample procedures have been briefly reviewed and one given in complete detail. Performance that can be measured quantitatively in terms of speed and accuracy appears on the surface to be an easy problem, but a few of the uncontrolled variables that may be carelessly overlooked have been called to the reader's attention. Such quantitative measurements must be set up with great care. Even more chances of error come into the qualitative evaluation of aspects of performance that do not lend themselves to quantitative measurement. Such qualitative appraisals are essential on many jobs and in many academic subjects in the classroom and laboratory. Methods of planning these appraisals need improvement if rating is to have any satisfactory degree of objectivity. It is hoped that the illustrations here that have achieved some degree

of success will be helpful to teachers and personnel technicians facing similar problems in their respective fields.

Methods of Rating. If the work sample has been carefully selected and the handling of the situation carefully planned in advance, the rating problem has already been much facilitated. The next step, then, is to define the rating factors. Roughly, there are two categories of rating factors to be considered—quantitative and qualitative. To a certain extent these may overlap. However, some strictly quantitative measures include the following. Accuracy will frequently be important. The squareness of a woodworker's joint can be checked to perhaps one-tenth of a degree. Errors in typing or in computation can be counted and tabulated. Speed is measurable in terms of words typed per minute, time taken to finish a welded joint, number of operations finished on a bookkeeping machine during an hour, etc. In other words, the test can be designed to show how many finished units of work product can be put out in a given time or to find how long a time is required to finish one unit of work product. A third category might include various measurable characteristics of the final product for which special equipment is required. For example, in electrical work, resistance can be measured, and this is said to be a very good indicator of how well a soldered connection has been done. Strength of a welded joint can also be determined exactly with instruments.

The second broad group of factors covers those which cannot be given an exact figure through the use of instruments. In this group the greatest differences of opinion are likely to occur. On a very rough scale of three or four levels of excellence, perhaps the surface finish of a woodworker's product could be rated by several experts with some degree of uniformity, but greater discrimination on the basis of more than five levels would probably be too unreliable to be worth while. Rating on work methods will probably furnish occasion for radical disagreement, especially where several quite different methods can be made to result in equally excellent final products. No one would dispute the merits of boring a hole nearly as deep as a screw is long before driving the screw into hard wood, but there are many other operations not generally accepted as essential in the trades. For instance, mechanics in some shops are expected to check certain steps with measuring instruments every time,

while foremen in other shops would regard such checking as an unnecessary waste of time.

The evaluation of complex skills like counseling is by far the most qualitative process conceivable, but the raters are more likely to be objective in their approach than skilled tradesmen, since their scientific orientation would presumably enable them to detect sources of unreliability that would not be evident to most shop foremen. The sample given in Appendix A is a rather involved rating problem, but it can be handled with a fair degree of reliability with the scale presented below.

Using the scale in Appendix C, three experienced psychologists who had received training and practice in nondirective counseling and psychotherapy rated the nine student counselors whose responses appear in Appendix B. These ratings were done independently by each psychologist; then a discussion was attempted in an effort to find reasons for differences in ranking of the students by the three raters. Very little of importance came out of the discussion. Differences in emphasis upon various errors or different strong points of the counselors seemed to cause disagreement in rating.

The results are summarized in the table below. The rank difference correlation method was used. Since the sample is very small, the findings are tentative rather than conclusive.

<i>Student code number</i>	<i>Counselors' rankings</i>		
	A	B	C
1	2	2	5
2	9	8	9
3	5	4	3
4	8	7	8
5	1	1	1
6	3	3	2
7	6	5	4
8	4	6	7
9	7	9	6

Rank order correlation coefficient: AB .90, BC .80, AC .78.

The three psychologists are designated by the letters A, B, and C, which appear at the top of the columns at the right, while the students are coded by number in the left-hand column. In such a subjective activity as counseling, the ratings might be expected to show much less agreement than they do here. These high correlations

are doubtless accounted for in part by the similar orientation of the raters in the nondirective approach, at least so far as this particular problem was concerned. Without such a definition of approach at the outset for both students and raters, much greater variability in ranking would have been anticipated. Had the identity of the student not been concealed by a code number, other factors would have entered the picture, such as effectiveness of use of the voice, facial expression, gestures, general appearance, etc. The amount of disagreement that would have resulted from the introduction of these judgments is a matter of speculation.

The fairness of a performance test for counselors with these factors eliminated is a controversial issue. When well-chosen statements are given in a hesitant manner with inappropriate facial expression and other intangible distractions that are expressions of an ineffective counselor personality, the effectiveness of what is said is without doubt greatly impaired. There are important intangibles in the counselor's manner that have nothing to do with his choice of words. These variables often operate to increase the effectiveness of the counselor's work, but they may also impair his efficiency to a marked degree. Personal likes and dislikes of the rater will unconsciously enter the picture here in spite of every effort to disregard them. On most candidates for jobs, several raters will probably agree upon their most significant observations, but occasionally they will differ radically on one and not know why. Thus the three psychologists may have achieved a measure of objectivity at the expense of completeness and fairness. Other schools of thought besides the nondirective may be regarded as having some merit. If they deserve a place in approaching the problem presented, the student counselor may be unjustifiably limited by the instructions to use exclusively nondirective procedures.

The above criticisms of the research presented here are intended to remind the reader that oversimplification of performance tests may have undesirable effects in lowering validity. There are limits to the usefulness of strict definitions of concepts that must be kept in mind in planning a rating procedure.

This chapter has covered a few main ideas regarding selection and use of materials, rating methods, etc. Where approaches to a

skill are radically different depending upon the traditions of different schools, credit must be given liberally for either acceptable approach, unless, of course, the instructions clearly define the method to be used. Even with such definition, the test may be considered unfair by advocates of the opposing technique. Thus there are many angles to consider in planning performance tests.

REVIEW AND TRYOUT OF A TEST

AFTER much concentration on an item, revising it in various ways, the test writer is likely to become so involved in the problem that he is unable to stand off and look at it from the point of view of the examinee reading it for the first time. He is unable to see it in the proper perspective. Having originated the idea, the test writer is quite sure what it means to him, but the reader, even if he is fully acquainted with the subject matter, will not necessarily think in the same channels. The writer, after hours of labor on his test, may be in a rut in his thinking. He may momentarily be unable to see that any of his material could possibly have any meaning other than that which he originally intended. The necessity of review is therefore obvious.

How thorough the review should be may depend upon a number of variables. For academic use in the classroom, the test writer may not find it practical to do more than look over the items himself, preferably at least one day after they have been prepared. The lapse of time between preparation and review may serve to help him out of the rut he has inevitably slid into while writing. But there is still considerable chance that he will miss something important, no matter how expert he may be. If the material is to be standardized ultimately, or if its use on a large scale is contemplated, review by someone else is advisable. Such a checkup may be for the purpose of evaluating content only, or it may be merely an editorial review to pick up errors in the mechanics of English usage, grammar, spelling, punctuation, etc. The latter may seem trivial and unimportant, yet numerous illustrations can be given where a misplaced "only" or an omitted comma has a detrimental effect upon the intended sense of the sentence. Perhaps the most logical sequence for the various types of review, if all of them are to be used, is (1) to begin

with the content or subject matter, then (2) to proceed with technical points of construction, and finally (3) to make sure that mechanical details of language are accurate.

Subject-matter Consultants. Not every expert in a given field would be useful as a consultant in a testing program. A professor of engineering in a university might be too academic and theoretical to be able best to evaluate examination material intended for use in the selection of engineers for a state highway department. This would not necessarily be true, of course, if the professor had had practical experience outside the teaching field, but one should be on guard, in public personnel work, against making tests too "textbookish" in content and wording. A mechanical engineer might be efficient in his field, and he might very well know what is important in repairing trucks and heavy equipment, yet he could turn out to be a very poor consultant on a personnel-selection test for automotive mechanics. He might never recognize that the candidates might be less able to think verbally in technical terms than he could himself and at the same time be able to perform the duties of their jobs, which call for a minimum of comprehension of written matter. Illiterate men have often been known to succeed to a rather outstanding degree in some manual occupations. One such highway-equipment mechanic seen by the writer was scarcely able to read at all, but when told how to service a new piece of equipment, he never forgot. The foreman gave him directions only once, and every detail was thereafter carried out to perfection. To eliminate such a competitor on the basis of comprehension of long, involved sentences prepared by an expert in the form of a written test would destroy the validity of the selection procedure.

To write good test items in a field not known thoroughly by the test technician may require the combined efforts of one trained in methods outlined in this volume and a specialist in the subject matter of the examination. The technician cannot feel safe in pulling his material out of textbooks, standard practice manuals, handbooks, etc. On the other hand, the specialist who modifies some of the unsuitable questions taken from printed material or who feels compelled to compose completely new ones may employ double negatives, specific determiners of all sorts, improper distractors, etc., unless the technician watches for these errors. Therefore good team-

work will be essential if the final product is to have any substantial value.

The writer had greatest difficulty with consultants in manual occupations. In order to avoid the overtechnical approach, he attempted to use men on the jobs in question and foremen supervising these men. The instructions for the review had to be very simple. The consultants were directed in part as follows: "Look at this test and the answers to the questions there. Tell me if any of these wouldn't make sense to mechanics, and we might be able to change them a little so that they would make sense." One mechanic labored slowly through four questions reading aloud and then said, "They ain't gonna know what you mean by this'n." The writer asked why, but could get nothing more than, "They ain't got none of them on them Chevrolet trucks."

Lack of verbal facility made it almost impossible for many of these consultants directly to offer constructive criticism. Composing new material had to be done by simply asking the consultant a few easy questions about the work. The technician would then formulate a problem, for example, for a multiple-choice pattern. He would write down the correct method, tool, or material as stated by the consultant. Then he would ask the consultant what wrong methods, materials, or tools might be used by workers not satisfactory to him. From this the wrong choices would usually result. However, the appraisal of the final product by the consultant was uncertain; he would read it, assume a puzzled expression, and finally mutter something like, "I reckon it's all right."

Fortunately, really helpful consultants can be found in nearly any field if the technician investigates far enough. Trade school instructors appeared to the writer as the best for skilled occupations, since they usually could verbalize at least fairly well and often understood the learner's problems better than did the foremen. In the more technical business and professional fields, several consultants were found who proved to be enthusiastic, efficient, and objective. Cooperation with them was decidedly effective.

Some personnel agencies have attempted to solve the consultant problem by training their own specialists in principles of test construction so that their agencies could actually write their own test items and tests. This procedure has met with a considerable degree

of success. The results obtained in the form of examinations for accountants, engineers, nurses, social workers, etc., have usually been of a high quality, as demonstrated later after a tryout. A few of the worst examples seen by the writer, however, have convinced him that there are pitfalls in this method if the specialist is held finally responsible for his items without review by one more experienced in examining. Not all outstanding specialists in any field (even psychology) can learn to become good item writers. The reasons for this observation are not entirely clear, but it is a matter of common knowledge among graduate students in psychology that some well-informed experimentalists in the field fail miserably to make themselves clear on examination questions. Preparation of this kind of material for measurement of achievement or aptitude in any subject is a highly specialized skill in and of itself. It requires aptitudes for clear thinking and self-expression as well as time to learn.

If a consultant is to be hired on a temporary basis rather than in a full-time capacity as an examiner, the instructions given to him are important. He must understand exactly what the technician wants him to do with the material. While for some relatively less educated individuals (in the formal sense) the task could be explained in terms such as those quoted in the example of the mechanic above, this simple formulation would not fit other circumstances. If the test technician is not expecting to be present all the time the consultant is working with the material, it will be essential to safeguard its security in every way possible. The specialist may or may not think of this point himself, and nothing will be lost by reminding him of it. He will probably be glad to follow the suggestion that the papers be kept in a locked file or desk, and the request that the content not be discussed with anyone except the examiner cooperating with him in its preparation.

When these points have been clarified orally or in writing, the next step is to give the expert a frame of reference ¹ for his evaluation of the content. This is best done by mentioning specifically and, if necessary, by describing the class or classes of positions for which the selection procedure is intended. Then the consultant can judge

¹ A frame of reference is a statement or description of the conditions or situations under which and for which the test or item will be used. It may be closely parallel to the word *Konjunktur* as used in economics or other social sciences.

whether the questions are suitable or not by imagining that he is a candidate for a specific job as he takes the test. In other words, he is not expected to decide whether the material is entirely accurate and stop there. He must have its purpose in mind. The wording of this part of the instructions will vary somewhat with the background of the specialist whose services are used.

To determine whether the subject matter of items has any application to duties of the job is not sufficient. The consultant must also decide whether he thinks good candidates are likely to get the right answer, while poorly qualified ones are likely to get it wrong. This is a nontechnical way of asking his opinion concerning the validity of each item. He may want to supplement a negative answer to this question with suggestions that will improve discrimination value.

The above question logically leads to two others which may often have to be investigated before validity can be improved. The first of these is whether there is one and only one correct or best answer. The importance of this has been stressed earlier in this volume several times. At this point the emphasis should be added that often it is the subject-matter reviewer who discovers for the first time that there is no entirely correct answer in a multiple-choice item as it stands. Some inaccuracy or exception overlooked by the originator of the item may invalidate it entirely. Equally often in personnel experience, the subject-matter reviewer discovers an item which has not one but two possible and equally good answers. Again the rut mentioned earlier prevents the writer of the original from thinking of an alternative interpretation which must be accepted as entirely justified.

The second question that follows from a consideration of subjective validity is the plausibility of wrong choices in the multiple-choice form. This can be explained to the reviewer in some such way as the following: Are these wrong answers so obvious that even poorly qualified people would seldom if ever choose them? Could they be made more nearly correct and still be inferior to the right one so that good candidates would recognize them as wrong, while poor ones would not? Useful suggestions should result from these inquiries if the reviewer has superior verbal facility. A modification of the same consideration could apply to false statements in true-false formulations. The one difference would be that an absolute criterion

of truth or falsehood would apply, whereas in multiple-choice form it is acceptable practice to call for the best of several possibilities, even though none of them represents an ideal or perfect solution.

Finally the critic is concerned with matters of accuracy which he is in the best position to judge. The first of these could be the wording. Every profession has its jargon; whether we like them or not, we must accept technical terms in general use, even though we may think the synonyms we have been using are identical in meaning and therefore just as acceptable. A sociologist who is dealing with data regarding a family may use words that are not acceptable to the professionally trained social worker. The latter may be reviewing an item originated by the former, and he may object, not to the content, but to the terms employed. If the question is intended for social workers, it must be in their terms. It must be expressed so that they will not have any misunderstanding of the language, provided that they are efficient in their work. The subject-matter reviewer is the one best able to supply the correct terms for the trade or profession. If a controversy is involved among different schools of thought in a field of knowledge, the critic is the one best able to decide which terms will be understood by every well-qualified examinee without definition and which ones will need to be defined to make the question clear. In education, philosophy, psychology, and psychiatry, the necessity for such clarification of terms is usually obvious. The word *instinct* or the word *attitude* could be extremely ambiguous, for example, unless defined, since each has been used by different well-known writers in different contexts with meanings that are far from identical. Specific unlearned behavior patterns, like nest building in birds, and vague tendencies, such as gregariousness in man, have both been called "instincts," but they are far from alike. Context seldom gives the reader enough of a clue.

The critic's second consideration in regard to accuracy may already have been covered in an earlier step. This is the accuracy of keyed answers, which may have been considered with the question as to whether there could be more than one right response. Routine as this point may seem to the inexperienced examiner, the technician will often discover errors in the scoring key that seem utterly absurd. These frequently appear after scoring is done and when papers are being reviewed by candidates. Inability to defend the rating in

such a situation is often quite embarrassing. The source of the error may be rather obscure. The technician may have pulled the question out of a book or manual. He may not have been too well acquainted with the point in the item and may have carelessly made a clerical error in keying it *true* when it should have been *false*, or D when it ought to have been B. The expert, who is familiar with the area of subject matter covered, may be quite likely to catch this kind of error if instructed to do so.

After the specific items have been subjected to the above critical analysis, the subject-matter consultant will probably be in a position to comment upon the selection of the sample of material. A good outline of content, if available, will be helpful at this point. The consultant may then be able to suggest important areas, if any, that have been omitted entirely or minor details that have received overemphasis on the examination. Changes can be made if needed to make the sample more representative of the duties. General comments concerning the forms of questions used or the organization of the material may occur to the critic at this time, and he should be encouraged to give these.

One common fault of subject-matter consultants is blanket approval of the technician's work without assurance that he has actually gone into detail in his evaluations. The technician may like to have his ego inflated a little by being told that he has assembled a good examination, but he does not want to have it deflated again later because his reviewer has overlooked important points that need improvement. There is always the possibility under such circumstances that the critic is sincere in believing that no further work needs to be done on even the most minor detail of the material, but the technician would like to make sure that the criticism has not been too superficial. In some instances, the only answer to this problem is to secure the services of a second reviewer if time permits. If two reviewers agree on what is good about the test and what needs changing, especially if the two have worked on it independently, the evidence is more conclusive than if only one expert was consulted. What if two experts disagree? Difference of opinion may even be radical regarding a single question, in which case it may be well to omit it entirely, since it could hardly be defended under such conditions if both experts were recognized as capable. The value of

an entire test is much less often disputed by two critics, but a difference of opinion may occur, for any of several reasons. An important one has been observed to be rooted in prejudice against some type of item, such as multiple-choice or true-false. The prejudice may extend even to the whole idea of tests in general, and in such cases consultant services are not likely to be effective. In the final analysis, the test technician will probably have to evaluate the expert criticism he has received and determine whether it is at all adequate or whether he will need somebody else.

An example of the suggestions offered by two consultants on an item might be of interest here. An examination for office supervisors was being prepared by a public personnel agency. Two consultants with extensive experience in business administration were secured at two different times to evaluate the material. One of these critics was somewhat misinformed on the potentialities of objective tests and insisted that aptitude for self-expression and quick decisions could not be appraised by such a test so well as by an essay form or perhaps an interview. His criticisms were, on the whole, constructive, however. The other critic favored the objective-test procedure and gave detailed comments that helped to ensure complete accuracy.

The item in question appears below:

If two efficient employees were constantly quarreling about details of their work to the extent that their conduct was lowering morale in the office, the supervisor would probably get best results if he

- a. fired both employees
- b. called both into his office together for a conference to discuss the matter
- c. fired one of the employees
- d. called each one into his office separately for discussion of the matter
- e. reprimanded both employees in the presence of their fellow workers.

The intended answer in the original key was d, but the second critic (who favored the objective test) pointed out that choice b could probably not be defended as wrong. Since two correct choices would not be advisable, he suggested that b be eliminated and that the choices be relettered. Then, of the four choices, there would be only one correct answer. The first reviewer (who did not favor objective tests) maintained that the item was a very poor one as it stood, since the best answer was not even there: to assign the two

workers to different divisions of the organization in order that they would have a minimum of contact with each other. He further maintained that a free-answer form or even an oral interview question would be far superior, since it would allow credit for constructive thinking through two stages: the conference, and if that did not work, the separation. The test technician then raised the question whether the same objective could be partially achieved by adding the word *first* to the premise and revising the item to read as follows:

If two efficient employees were constantly quarreling about details of their work to the extent that their conduct was lowering morale in the office, the supervisor would probably get best results if he first

- a. fired both employees
- b. fired one of the employees
- c. called each into his office separately for private discussion of the matter
- d. reprimanded both employees severely in the presence of their fellow workers.

Both critics ultimately agreed that the correct answer to the revision was c, but the first one still preferred essay or oral form, while the second regarded the revised edition as entirely satisfactory. Both had submitted valuable ideas not taken into account by the originator of the question, and the originator was able to incorporate both suggestions into a satisfactory revision. Ultimately, the oral form of the question was introduced into interview parts of several examinations for supervisory positions with discriminating results. Item analysis proved the good discriminating quality of the multiple-choice form as well. This was an exceptionally smooth-running situation. Not all reviewing problems work themselves out as simply as this one did. Expert opinions in any subject-matter area need to be verified through tryout and item analysis. They are subjective and by no means infallible.

Technical Review from the Construction Viewpoint. This topic overlaps the previous section of the chapter, since content and construction go hand in hand, but for the most part the construction aspect is a technical function that is the ultimate responsibility of the examiner rather than of the subject-matter specialist. In an aca-

demical situation, the review for construction principles is usually integrated with that for content and is done by the same person.

This review brings into application many principles discussed in detail earlier in this volume. Specific determiners of various sorts are most likely to escape the attention of the writer when he is working on the original version of his test. These are apt to stand out sharply the second time he goes over the material. He may find that he has unconsciously made the correct choice the longest almost every time in multiple-choice patterns. He may even have made the right one consistently the shortest. He may have unconsciously followed fairly uniformly a stereotyped pattern of key answers, like T, T, F, F, T, T, F, F or B, D, A, C, B, D, A, C. Shuffling the numbers of the items before the rough draft is copied in final form may be all that is necessary. Double negatives are also particularly likely to have escaped his notice the first time. Errors may have crept into the format or arrangement of choices as they were typed. Of course the mechanics of grammar and spelling as well as punctuation will require editorial review. Confusion in the interpretation of questions caused by omissions, misspellings, and careless typographical errors often lead, whether justly or not, to accusations of unfairness of the examination. The writer recalls several instances of such carelessness in overlooking mechanical errors that much confusion and ill feeling were caused. One such test, a final examination in a large university class of 60 students, was full of omitted words, phrases, or actual entire choices. The choices were often labeled A, B, C, C, or A, B, D, D. Sometimes the correct choice itself was not even there. The test had presumably been proofread by two assistants in the department, but the errors were discovered by students and instructor while the test was being taken. The instructor had left the matter of mechanical errors entirely to these assistants after assuring himself that his original copy was correct. He had not had time to check the mimeographed material in detail before examination time. As the errors were discovered and questions raised, confusion increased, mumbling conversation began here and there, and signs of cheating beyond the usual expected amount were evident. Much time was lost for everyone concerned, and the measurement was practically invalidated. In another case, the errors were discovered by the

instructor on his original as well as the copies after scoring had been done. Much time was consumed in rescoring, and examinees were bitter about their ratings to a degree that even the corrections could not entirely overcome. Usually such errors, if serious, count against the student or candidate for a job rather than for him. If people are to have respect for tests, they must be accurate and fair. Mere carelessness will destroy good rapport in either academic or personnel settings.

The best way to proof anything, assuming that the original copy is correct, is to have two persons work on it. One can read the original while the other checks the stencil. Legibility assumes some importance also. A dimly printed copy is a strain on eyes and mind. An examinee has enough problems even if the material is all there. He should not have to decipher or guess at anything that is only half printed. Thus we see that even such details as inking the stencil properly assume importance if confusion and loss of time are to be avoided. Every detail of test preparation, administration, and scoring well done wins good will for accurate, scientific measurement in the long run.

Item Analysis. After the extensive reviewing procedures just described have been carried out, the inexperienced student of mental measurements is likely to conclude that the carefully prepared final product now surely bears a guarantee of at least something close to perfection. Unfortunately, however, experience has taught the more sophisticated educator and personnel technician that the first actual tryout may show up many unrecognized defects.

If the initial population is a small one, nothing much beyond crude inspection techniques can be applied. What is observed by these methods may be far from conclusive. Yet there are a few ideas which may be useful in inspecting the results from the first small sample of subjects who take the test, and these will be considered first before going into the statistical procedures for use on large groups.

The first question to come up is one of difficulty level. If an item is passed by 95 per cent of the group or better, it is obviously an easy item which does not discriminate very well except, perhaps, at the very bottom of the scale of that trait or group of traits being measured. In other words, it may be discriminating only between

failures and miserable failures, neither of whom would make good students or employees. Such an item might do as a warming-up process at the very beginning of the test to help the examinee to recover from some of his apprehension and stage fright. But very many such easy tasks would be a waste of testing time that would accomplish no purpose. Similarly, a task so difficult that 5 per cent or less were able to succeed at it would discriminate, if at all, only between superior and outstanding students or employees, and indeed, success might be a mere matter of chance or of the possession of a bit of useless detail. If most of the test consisted of this kind of situation, it probably would be a rather useless measuring instrument as a whole. Perhaps the distribution of item difficulty should ideally conform somewhat to the normal curve, most items being about middle level and a few at each extreme. The smaller the group, the more irregular would be the curve of difficulty and the less reliable would be the discrimination obtained by the test.

In item analysis, the second problem that arises is the discrimination value of each item. By this is meant the degree to which the single item separates the superior from the inferior individuals in the trait or group of traits being measured. Ideally each item should then be correlated with a true evaluation of the trait or traits in question—some criterion other than the test itself. Perhaps the point can be clarified by an illustration. If a test for selecting salesmen were to correlate moderately highly with actual success on the job in selling, then it would be helpful to know how much each test item contributes to the prediction of job success. Efficiency in selling would then be called an external criterion, outside of the test itself. Such an external criterion would be difficult to establish, but this must ultimately be done if the validity of the test is ever to be established.

Before an external criterion of any sort is available, however, there is a third question that will probably be much easier to answer than the second: To what extent does each item measure the same variable or group of variables that the test as a whole measures? If the initial tryout of the test has been on a large enough group, perhaps 200 or more, an internal criterion may be used. In other words, each item may be correlated with the test as a whole to determine to what degree success on the item is related to success on

the entire test. If the measure as a whole is not valid, this process may be useless, but some improvement may be brought about by the study of what is known as the internal consistency of a test. When items are found which contribute nothing, they should be omitted or rewritten. If as many good students (in terms of total score) fail an item as there are poor students who fail it, the item contributes nothing to the intended purpose of measurement.

Sometimes an item may discriminate decidedly in the wrong direction. Most superior applicants for the job miss it, while most low-scoring ones get it right. If this happens, the item may be one which if answered by mere superficial thought is easy enough, but if analyzed more thoroughly becomes ambiguous, confusing, or very difficult. To include several such items would lower both reliability and validity of the entire measurement.

Questions that appear valid to their author may turn out to depend too much on chance. They may fail to discriminate at all for reasons that are anything but clear. Subjective judgment, even when based on long experience, is not infallible by any means, and item analysis is considered essential by the best authorities in test development.

Various item-test correlation techniques are in current use. Some of these are quite involved, though extremely accurate. These meet better the requirements of research projects than those of the usual practical situation in which time is important. A few simpler methods are in common use, and one of these will be described in this chapter.

If 100 subjects are available, they may be divided into two groups of 50 on the basis of total score. The median, or middle score in the distribution, can be taken as the dividing point, except that occasionally several individuals make this score. In such a case the division will have to be rather arbitrary, since some with this score must fall in the upper group and some in the lower in order to make the two groups exactly equal. An even number of cases must be used in order to have upper and lower groups the same size, and fewer than 50 in each would make the results too unreliable to make the process worth while. Two hundred cases would be almost ideal, since the ease of dealing with numbers of cases exactly the same as percentages within each group would have a decided advantage. Numbers greater than 200 would increase reliability somewhat, but would increase the complexity of computing percentages.

After the division is made, the next step is to count the number in the upper 100 cases (if 200 are used) who chose the correct answer. This is done for each item in the test by going through the papers and tallying for each question separately. The clerical work involved at this stage is tedious, but unless the test is machine-scored, there is no short-cut method. Machine-scored tests may be tallied in this way easily by use of the graphic item-counter attachment that can be obtained with the I.B.M. test-scoring machine. Though difficult to learn to set up properly, this remarkable device saves enormous amounts of time in large-scale test-development operations. It rapidly prints a column graph to show the number of correct answers to each item on an entire test on one sheet.

If the item analysis is to show not only item difficulty and correlation of each question with total score but also the value of wrong choices in a multiple-choice form, an additional step is necessary. This consists of tallying the number who selected each wrong choice. An answer which was intended by the writer as wrong, yet plausible, may not have been chosen by anyone in the high group and by very few in the low group. Such incorrect answers might just as well have been left out. If only two of the four answers to a question were marked by anyone, the right one and one wrong one, the other two being ignored entirely, the chances of success by mere guessing are not one in four, but one in two. Of course there is some value in developing a technique in writing good distractors (wrong answers), but the prediction that many will select a given distractor often turns out to be incorrect. This step will then enable the examiner to improve many poor distractors before repeating the test.

Logically, the next stage of item analysis will be to repeat the same counting of right and wrong answers within the low group. This half of the cases will be likely to select wrong answers more often than did the other half. They will more frequently omit a question entirely, and this must be tallied, too. Their right responses will be less frequent, on the whole. Yet now and then a noteworthy exception will occur. These cases will call for further study.

As examples will be given the data on one very good question and then on one very poor item. The population in both cases consisted of applicants for entrance-level clerical positions. None failed to attempt an answer.

No. 44	Choice A	Choice B (correct)	Choice C	Choice D
Upper 100	14	58	16	12
Lower 100	32	36	21	11

From inspection, the results seem obviously favorable. The high group appears to have answered it significantly better than the low group, showing that it discriminates in the same direction as the entire examination does. The wrong choices were not too obvious, and were selected by a sufficient number in each group.

The data below seem equally convincing in the opposite direction:

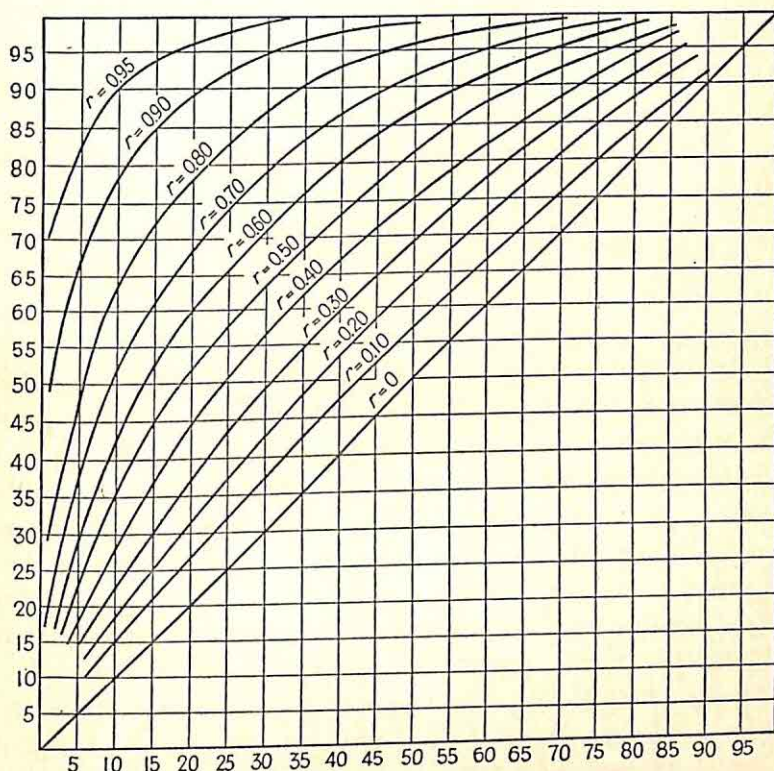
No. 57	Choice A	Choice B	Choice C	Choice D (correct)
Upper 100	46	6	3	45
Lower 100	21	0	10	69

In the above data, there is some evidence that the key is incorrect or that there may be two answers with merit, A and D. Choice B is of doubtful value, and probably seems too unlikely even to the poorest applicants. Choice C is not much better. Perhaps the entire question should be omitted and a better one found. If A is incorrect, it should discriminate in the opposite direction by being selected more often by the lower group than by the upper. If D is right, more of the upper group should mark it.

After these tabulations have been completed and inspected, there is still another part of the process that will make the findings more meaningful. This is the computation of the item-test correlation coefficient, which may be done from the chart on page 157. For item 44, the problem is solved as follows: First find 58, the number in the upper group passing it, at the left. This would fall just a little below the horizontal line for 60 per cent. Follow this line across the chart to 36, just beyond the vertical line for 35. This 36 represents the percentage of the lower group that found the right answer. Place the point of a sharp pencil just below the 60 horizontal line and slightly to the right of the 35 vertical line. This estimated point will fall roughly about midway between the .30 curve and the .40 curve, resulting in an interpolated value of about .35. Greater accuracy of measurement of space between these various lines would probably have little meaning in connection with this problem.

Although the writer does not know of a well-established criterion

value for an item-test correlation below which a question would be considered too poor to be retained, the coefficient of .35 found for the above item is suggestive of a fairly good question that is definitely discriminating in the same direction as is the test as a whole. Therefore item 44 is regarded as satisfactory.



Computing chart for tetrachoric correlations with the criterion dichotomized at the median. (Adapted from Mosier, C. I., and McQuitty, John V., "Methods of Item Validation and Abacs for Item-test Correlation and Critical Ratio of Upper-Lower Difference," *Psychometrika*, 1940, 5(No. 1):57-65.)

Number 57 yielded 45 successes in the upper group as against 69 in the lower. The 45 horizontal line meets the 70 vertical line on the chart far below the diagonal representing a correlation of zero; therefore the value of r (the correlation coefficient) would be a minus quantity. Its exact value would be unimportant, since the item obviously discriminates in the wrong direction and should be rejected or revised completely.

Very low correlations of .15 or .10, even though in a positive direction, would make the material questionable. Further investigation beyond the mere value of the tetrachoric r would be advisable in such cases. Unless the wrong answers were decidedly poor, the material could be retained without damage to the test as a whole, yet it would contribute little. High tetrachoric correlations (.85 or better) are rare in most tests.

The stability of values found on this chart is an interesting problem for speculation and, indeed, for research. Only one repetition of such an item analysis on a group of equal size has been attempted by the writer, and there were a few marked changes from the first trial of the test to the second. Most of the r values stayed the same or nearly the same in both trials, however. Thus the process seems decidedly worth while in screening out poor material. The difficult point occurs when the examiner tries to reason out what made a particular item poor. The defects in its construction may often be rather obscure, and the statistical process may not tell the examiner much about how to improve some of his material. Trial and error over several administrations may ultimately be the only way to perfect a test in every detail, but far too often the statistically unsatisfactory questions are violations of the rules of construction discussed earlier in this volume. Following these rules carefully may avoid much correction later, but in no case will technical skill in construction replace the tryout and item analysis.

VALIDITY, RELIABILITY, AND STANDARDIZATION

THE average high school teacher or college professor, unless he has specialized in mathematics, is likely to have a limited knowledge of statistics. The same is likely to be true of many personnel managers and technicians in business and governmental agencies. Probably many such persons have had at one time or another a course in statistics applied to psychology, education, or business problems. If this course was as condensed as many of them are, the student probably had a good case of academic indigestion before it was over. For all but perhaps the most brilliant students, statistics requires time to absorb. Such concepts as the standard deviation, for example, cannot be explained very convincingly in a few sentences. Only through use of it in solving problems in which the student is interested will it come to have much real meaning. The relatively new developments in factor analysis may be compared in some respects with Einstein's theory of relativity. Perhaps only half a dozen men in the world really understand them thoroughly. Leading research workers in test development have spent years studying and using statistical tools, and they will stress the need for these tools for every investigator in the field of testing.

However, lack of mathematical aptitude or interest, together with some degree of emotional blocking whenever figures of any complexity are encountered, probably explains why the majority of teachers and personnel technicians are reluctant to go very far with statistics. Disuse has resulted in much forgetting. Many of them would like to construct better tests if not too much mathematics were involved.

To give a course in statistics is outside the scope of this volume,

and the reader is referred to the excellent treatments of the subject by Garrett, Guilford, and others, listed in the Collateral Readings at the end of this book. This chapter will be concerned only with some of the general principles which govern the usefulness of a test once it has been well constructed. Only its practical value can demonstrate that it really has been skillfully planned in every detail. Various aspects of its practical applicability will be considered here. At least a few elementary procedures will be suggested that the reader can carry out with a minimum of statistical background, if he has access to enough subjects.

The Criterion. Validity is defined as the extent to which a measuring instrument measures what it was intended to measure. If through ignorance a person should construct a barometer with the purpose in mind of measuring atmospheric temperature, he would probably find out that his instrument was not valid. It would be measuring atmospheric pressure instead. Similarly, a test may be constructed to measure achievement in a given occupational setting, such as highway-maintenance foreman. Yet the questions may be so worded that they measure vocabulary rather than common sense in supervising construction laborers. Suppose, for instance, that the test consisted of items like this: "What methods would you find most useful in orienting new employees under your supervision?" For office supervisors, this sort of question might be valid, but for a foreman of limited academic education, it certainly would not. He might well know what to do, but he could not answer such a question, since he had never seen or heard two or three of the key words. A change to the wording, "How do you break in a new laborer on the job?" would make a great difference. A test of items in job language may have face validity (seem to apply to the work) and also actual validity.

Thus we have shown that validity must be defined in terms of the purpose for which the test is to be used. To say that a personality inventory has been found to be valid is hardly specific enough. It might not be valid below eighth-grade level, since the items might be ambiguous to the less mature child but very definite in meaning to older subjects. Having observed this precaution in interpreting data, the investigator is now ready to consider the techniques in establishing validity. The entire concept discussed here is one that

has been neglected or passed over lightly for years. Only recently have investigators been much concerned about proving what has been assumed without question for so long. A test can more easily be shown to be reliable, and that has been considered sufficient by many research workers. Reliability is concerned merely with whether or not results are consistent. In the illustration of the barometer, the measurements might be quite dependable if the instrument was well constructed. That is to say, it could be a reliable measuring instrument, but still not valid.

To establish validity, then, it is first essential that an independent criterion be found. By independent criterion is meant some measure other than the test of the trait or group of traits which the test is intended to evaluate. Usually it amounts to some objective way of appraising actual performance on the job or in academic or in-service training. Some kinds of performance are quantitative to a sufficient degree that several raters following the same plan would arrive at the same numerical value for each sample of work evaluated. Not all kinds of performance can be thus measured, however, since subjective factors requiring qualitative judgment usually enter the picture to some degree.

Productivity of a typist can be stated exactly in terms of words per minute done with a given standard of accuracy over an extended period. Achievement in a training course or production on the job can be correlated with the test easily in such work. The criterion then amounts essentially to a large work sample which is correlated with a smaller one, the test. Qualitative judgment begins to play a part when attitude toward supervisors or fellow workers is taken into account as a part of success or failure. The test may be confined to speed and accuracy on plain copy. Susan may be slower and slightly less accurate than Jane on this, yet Susan is found to be particularly efficient on corrected copy. She always makes sure that her product is logical and makes good sense. If a much scribbled-over sentence does not seem coherent, she courteously asks her employer for help with it. She accepts criticism cheerfully, and tries to correct her faults. Jane, on the other hand, becomes confused if a few corrections have been inserted with a pencil. She dashes off a sequence of obvious nonsense and hands it in as her final product. She resents correction from her supervisor, and makes little effort

to improve herself. The test does not tell the whole story about these two girls. Supervisors rating them might easily differ in their evaluations of the acceptance of criticism and motivation for self-improvement.

Jobs or courses of training which seem to stress quantity and measurable quality of production are not so simple as they seem. The criterion may be limited to one aspect of work. Again, it may be inadequate to describe success or failure because likes and dislikes of personal characteristics enter in. Skilled trades, accounting, and machine operation are the kinds of work that have been shown to be best suited to validation studies, since criteria can approach objectivity, but serious errors can be made here. Motivation of the raters constitutes the most serious problem. Ratings of enlisted men by their officers are subject to many forms of inaccuracy. Yet intelligence tests have been correlated with such ratings in an effort to establish their validity. Which is worse, the criterion or the test? Ratings of welfare fieldworkers, family caseworkers, probation officers, policemen, teachers, salesmen, and nurses are sure to involve interaction of two or more personalities. An efficient employee may clash with his boss on one such job at every turn, while on another job he may get along smoothly with superiors and equals. No test of achievement in these occupations can be expected to yield high correlations with a criterion influenced by these variables. No test of aptitude can be expected to correlate highly with success in training for these occupations if personality interaction between individuals is allowed to enter to a considerable degree. For these jobs and many others, the problem is mainly one of defining an adequate criterion.

Two steps seem to be of paramount importance in defining the criterion for any educational program or job. The first of these is the construction of a good rating scale. The traits which are most significant are discussed and listed. Ordinarily, about 5 to 10 of these will emerge after discussion and perhaps a few trials. From this point on, authorities differ as to procedure, but the end product must in any case be free from ambiguous or vague items. One form of rating scale that has wide usefulness will be described here. Each trait is given a term, and below its name are given five descriptive phrases indicating various degrees of this trait. The rater is asked to mark the description that most nearly fits the employee or trainee. Em-

phasis is placed upon the phrase *most nearly*, since the rater may occasionally argue that no given answer exactly describes the employee in this respect. An example of such a rating form for one trait is given below:

INITIATIVE

- | | | |
|---|---|---|
| ____ Finds what needs
to be done and
does it. | ____ Usually goes
ahead on his
own. | ____ Works only if
given detailed
instructions. |
| ____ Needs constant su-
pervision. | ____ Must be constantly
prodded to get any-
thing done. | |

Sometimes the lowest degree of a trait is described first to avoid the halo effect (marking a person high straight down the line without much thought). The opposite, sometimes known as the horns effect, can in this way be reduced somewhat.

Once the rating scale has been carefully planned and subjected to a preliminary tryout, the research worker is ready for the second step. This consists in finding raters who are relatively free from personal bias and reasonably objective in their approach. Emphasis should be placed here upon *relatively*, since no supervisor can claim complete freedom from such influences, which often operate unconsciously in the individual. The examiner is forced to rely upon his own judgment in his evaluation of the motivation of raters. If a supervisor or administrator regards rating forms as merely required paper work to be dispensed with as quickly as possible and with the least effort, his subordinates will not be included in a validation study, since they will contribute nothing and probably spoil the results. The examiner can probably discover this indifferent attitude or the presence of strong prejudices through an interview with the supervisor. When strong personal likes and dislikes of the superior for subordinates seem to be evident, the results will be worthless. The selection of suitable raters may thus cover a large field of survey and consume much time. Only on large-scale testing programs will suitable conditions for a validation study be found.

When these two steps are completed, however, the way has been paved for proving or disproving the value of a test battery, or perhaps each part of the battery. The actual rating having been made,

the examiner is ready to apply the product-moment correlation technique to his numerical ratings and test scores. When this has been done, the resulting coefficients are often quite disappointingly low. If the examiner has any faith in the criterion which he has established, he can draw no other conclusion than that his test battery is decidedly in need of improvement. Much trial and error will probably follow before a reasonably satisfactory testing procedure is found. From these frequently demonstrated trends illustrated by military and industrial testing programs reported in psychological journals, it is plain that long years of research ordinarily are essential before claims of validity can be made for a new employee-selection process or system of educational measurements.

But the criterion may often be less satisfactory than the test. If the examiner cannot have confidence in his criterion, he has gained little, if anything. For some tests, the results have been correlated with those of another much more thorough test given to the same group that is assumed to be valid for measurement of the same trait or traits. The better-established test may indeed have been validated against a satisfactory measure of actual job success or success in training, and if this is true, the newer test should correlate highly with the older one. If it does, the evidence is at least fairly convincing regarding validity of the new test. If the recent measuring tool is in a relatively unexplored field, its author may have to fall back upon an internal criterion temporarily. For a time, all he may be able to establish is that each item in the booklet (or nearly all of them) can be shown to measure the same trait that the test as a whole measures. This item-test correlation method, described in the last chapter, is not to be considered final evidence of validity, since a test may be internally consistent throughout without measuring what its author thinks it measures.

The validity of teachers', officers', or on-the-job ratings would be difficult indeed to establish experimentally, and perhaps the examiner will not be able to do much but fall back upon logic to defend them. The literature is full of controversy and actual contradictions in results and conclusions along this line. Before these can be settled, much more research must be done in refinement of present methods. Some of the differences of opinion may have been the result of uncontrolled variables in the testing situation such as morale

or rapport. No satisfactory technique has been found for eliminating nonintellectual factors, such as examiner-examinee relationship or interaction, mood, emotional instability, and anxiety resulting in poor concentration. Measurement of aptitudes or achievement outside this framework of personality variables is impossible; therefore, they may as well be taken into account as a part of general mental functioning of the individual. If this level of functioning is variable in the test situation, it will probably follow the same pattern on the job or in training courses. Scores that are affected by these personality variables are realistic and more useful than would be pure measurements independent of such factors, were they possible. Therefore, the examiner may as well accept the limitations of his imperfect instruments until better ones are found.

Reliability. As stated earlier, reliability refers to accuracy. A measuring stick may be crude, so that in a general way it measures inches, but it may make many errors, probably usually of small magnitude. A poorly constructed scientific instrument does not give consistent results. It may measure roughly what it was intended to measure, but it will make mistakes, perhaps always in one direction, perhaps in any unpredictable direction. An unreliable aptitude or achievement test gives inconsistent results. If it is administered over and over again to the same group, great variability in the scores of most individuals will be found. If such a measurement is reliable, scores will tend to be quite stable, showing as a rule little variability from one administration to the next.

What makes scores stable or unstable? Statistics involved in answering this question would take a large volume to explain fully. It is recommended that the reader thoroughly acquaint himself with statistical tools through one or more of the excellent works on that subject. Only a few basic points will be mentioned briefly here. Other things being equal, length is probably the most important single factor determining reliability. A final examination of 10 multiple-choice questions in a course in almost any subject would cover too small a sample of the material to be fair to the student. Chance would be involved to a great degree in what points the questions required, and whether these were the points the individual student stressed most in studying for the examination. The results would therefore be unreliable. If another similar examination were

then given, the ranking in the class might well be radically different. A 20-item examination would be more reliable, but not twice as reliable. The greater the length, the more nearly perfect reliability will be achieved, provided that the material is equally good in quality. There comes a point, however, beyond which this increase in reliability becomes so small as to make any further lengthening decidedly not worth while. Fatigue and other variables begin to be increasingly important if the test is too long. Standard works on statistics give full explanations of the use of formulas for predicting increase in reliability with length of the test or the size of the sample.

The relation of test length to reliability is actually a rather complex statistical problem, and to give a brief summary of formulas used might be misleading. Adkins, Primoff, and McAdoo¹ give a concise review of these procedures, and more lengthy discussions may be found in standard texts on statistics. The examiner who plans extensive research with his tests is referred to these sources in order that he will not make a serious error in applying a tool that might not fit his particular problem. For most practical purposes, prediction of increase in reliability with increase in length may be of little importance, since the quality of material added may influence the actual reliability to an unknown degree.

Exactly how long a test or battery of tests should be is not an easy question to answer. For the measurement of such a relatively simple aptitude as pitch discrimination in the field of sound, 50 trials with various levels of difficulty, 5 to 10 trials at each level, would allow too much chance factor to enter into success or failure. Fluctuations of attention occur in normal subjects, and are beyond possibility of complete control. The effect of these may be less if a longer test is given, provided that it is not so long as to produce excessive fatigue. If a complex set of traits must be evaluated, a battery of several tests covering three or four hours, with appropriate rest periods between, may have to be planned. Thus the total number of items needed for vocational guidance, employee selection, or evaluation of educational progress may sometimes reach 600 or more. A typical written test for a one-semester or one-quarter course covering

¹ Adkins, Dorothy C., Primoff, Ernest S., McAdoo, Harold L., *et al.*, *Construction and Analysis of Achievement Tests*. Washington: Government Printing Office, 1947, pp. 155-160.

the entire text and lecture or other classwork would probably contain 100 to 200 objective questions. Some would be longer, but one as short as 50 would seldom be reliable enough to meet a satisfactory standard as to sample of material covered.

More important, probably, than mere length is the quality of the material. If the first 25 questions met all standards set forth in earlier chapters as to construction, and if at this point the writer ran out of good ideas and began to produce ambiguous, unimportant, poorly phrased items, the reliability of the 50-item measuring device would be lowered, not raised, by making it twice as long. The examiner who predicts increase in reliability with increase in length by statistical means must bear in mind the assumption that any additions made will be equal in quality to the original material. If this assumption is not correct, the statistical operations are worthless. Vague phrases that can be interpreted in a number of ways bring guessing into the situation. Chance factors and specific determiners enable subjects to answer correctly without the use of the mental function intended to be required. Clues that give away the answer (specific determiners) lower reliability, since only a limited number of individuals will lean on these crutches when they are present. Every double negative, every bit of excess verbiage, and every useless detail will contribute to lowered quality.

The computation of reliability may be accomplished in several ways. If a test is administered a second time to the same group, the correlation between the results of the two administrations may be computed. This is known as the test-retest method. It is time-consuming and often inconvenient to arrange for two meetings of the group, the equalization of motivation both times is a problem, and subjects when repeating the test may become bored unless the examiner is very skillful. They may also benefit from practice effects the second time, although the gain is usually inclined to be slight. If equivalent forms of the test are available, these alternate forms may be useful in place of a repetition of the identical material, but the assumption of equivalence of the two forms must be well defended.

If only one sitting of the group can conveniently be arranged, there are other methods which may be applicable. The simplest of these is the split-half method. It is accomplished, as a general rule, by correlating the number-right score of the odd-numbered items

with the number right of the even-numbered items. This way of splitting a test has some decided advantages over dividing it in the middle or any other way. If the first 50 items were matched against the last 50 of a total of 100, it is quite possible if not probable that the first half would be easier than the second half. Also, the subject matter or type of item might differ too much from one section to the other. If the odd-even division is used, the difficulty is likely to be more nearly the same. If there are several sections, such as vocabulary, number series, following directions, and analogies, each half would contain samples of each section, and the samples would be likely to be about equal in length and difficulty. The main advantage of split-half procedure is the time and effort saved. However, the resulting reliability coefficient represents the reliability of a test only half as long as the original. The test-retest coefficient is likely to be higher.

There are statistical devices for overcoming the disadvantages just mentioned. A correction can be made for the length of the half of a test, and only one administration of one form will be required. There are several Kuder-Richardson formulas discussed in the standard texts on statistics. These require different amounts and kinds of data about the test for their solution, but they are recognized as the best devices in test-development research, especially if complete item analysis has already been carried out. Full consideration of the applications of these various formulas would require several pages, and will be omitted here, but no test that claims to be well standardized would in modern times fail to publish data of some sort on reliability. Nearly all test manuals specify the method or methods used and give the results in some detail. Evidence on reliability is much easier to obtain and defend than that on validity. While validity has frequently been neglected entirely or passed over lightly, reliability is seldom omitted from consideration.

Results obtained by different investigators on different groups showing the reliability of the same test have revealed amazing variability. The instability of coefficients is probably less the outcome of different statistical formulas than it is the result of marked differences in motivation. Morale is usually the uncontrolled factor in industry and education. Intangibles that have to do with school spirit, interaction between teachers and pupils or management and subordinate

employees, group unity, competition, etc., all affect performance. No objective way of measuring or even accurately describing these variables is yet known; therefore, they are usually ignored or considered unimportant in writing reports of experimental studies. Authors often merely make assumptions as to the degree of cooperation obtained. They cannot defend these.

To obtain reliable test results, an examiner should see that several requirements are met. These can be summarized as follows: First he must construct or select a test that consists of a reasonable sample of what is to be measured. The absolute minimum cannot be stated in terms of number of items without regard for the complexity of the mental function, skill, or subject matter in question. A maximum number of items is again not an arbitrary figure, but one which is affected by complexity, fatigue after long steady work without a break, and the age range of those to be tested. Secondly, the examiner must consider quality of material. Anything violating important construction rules is apt to decrease reliability. Thirdly, he must take into account the variable of motivation, and this point is the most difficult one to control. Good rapport is essential here in a group test situation as well as in work with one individual at a time. The manner in which the testing program has been sold as a good idea, the purpose of it, and how it fits into other aspects of the job or of academic work may all make a difference in the cooperation which is obtained. Passive resistance to testing, shown in the form of coasting along with just enough effort to get by, is not at all uncommon. Yet actual research has been published in which the subjects were by no means trying their best. How many instances of this besides those known to the writer have appeared in print cannot be estimated. Many unreliable measurements have been made in which the tools have been blamed for failure. How many such failures would have succeeded with better motivation and with the same tools is a good problem for further investigation, but one requiring great skill. There are numerous such problems that would merely require repeating an old procedure under new and better conditions.

Norms. John Doe made a score of 70 in English usage. What does this rating mean? If the traditional school of thought is followed, it probably denotes 70 per cent of the items right. If 70 per cent is then defined as the passing point, again following tradition

rather than logic, John passed. Taken at its face value, this figure still has very little meaning when isolated from other data. The test writer may have thought that he was preparing a rather difficult examination, but after using it he may have found it to be much easier than he expected. Thus 70 might turn out to be the lowest score in a large group. On the other hand, it is quite possible that the originator of the examination gave little attention to the difficulty of his product, which resulted in only a few scores above 70 in a large group. The score of 70 thus comes to have entirely different significance under different conditions. Raw scores have no value for interpretation, even when reduced to per cent right. An arbitrary, traditional passing point or critical score cannot be applied under all circumstances. The criterion must be a useful one, one demonstrated to be adequate to separate successes from failures, one that makes sense.

Although it has been customary, for no good reason known to the writer, to establish a critical score at one standard deviation below the mean of the distribution, this arbitrary statistical practice has only the advantage that in an approximately normal curve it results in cutting off about 16 per cent of the cases on the lower end of the distribution. Whether all or most individuals above this score may be defined or predicted as probable successes must be demonstrated by experiment.

Establishing validity and setting up a true critical score go hand in hand. Before extensive data are available, only a subjective hunch can be used as a guide in setting up a passing point. Arbitrarily failing a predetermined percentage of a class or group of job applicants is the same as making a very unsafe assumption as to the quality of the group as a whole. Not all groups give a normal curve on any trait. Some are skewed to the right and some to the left, and some, indeed, are bimodal (having two distinct humps). Fifty per cent of one group may have to be failed if a barely satisfactory standard of job performance is to be maintained later. This would be particularly likely to occur when there is an oversupply of persons interested in a few jobs in one class. If there are many openings and few qualified applicants, the standard may have to be set quite low for practical reasons. Some groups may be quite well selected by entrance requirements for a course or job, so that very few test

failures should occur. Standards have to be flexible to meet conditions of supply and demand, but there are limits to how low the critical score can go without impairing the efficiency of the selection procedure, perhaps even to the point of making it entirely worthless.

If, because of the ease of explaining scores on a traditional basis, a civil service agency desires to establish 70 per cent as a passing point, the agency should avoid the wrong assumption that 70 per cent of the items answered correctly will always be a satisfactory criterion. A skewed distribution is frequently apt to occur. A normal curve could be approached closely, and still the rigid 70 per cent criterion could be impractical. For example, suppose that a raw score of 80 out of 140 items could be demonstrated to be a fairly adequate minimum standard of performance for prediction of success on the job. Also suppose that the highest score actually made in a large sample of the population was at or very near 140. The conversion to a per cent score would be very easy. A calculating machine could be set up in such a way as to make the conversion automatic. It would add .5 per cent to 70 for every point that the raw score exceeded 80. Thus 82 would equal 71 per cent; 139, 99.5 per cent; etc. Thus the range of 60 raw-score points was reduced to a range of 30 per cent. If the ratio of the raw-score range from the criterion to the highest earned or highest possible rating to the per cent range of 30 were more complicated, the computation involved would become tedious by hand, but it could still be done automatically by a calculator. Even in large-scale educational measurements, this reduction procedure might often prove valuable.

Statisticians have tried numerous ways of expressing scores. For research purposes, the use of some form of standard scores is often almost imperative. These standard scores are based upon the standard deviation, which is essentially a distance along the base line of a curve. This measure of dispersion, the most accurate one known, can be laid off three times along the base of the curve below the mean and reach about the lower end of the distribution. The same is true when it is laid off three times along the base line above the mean. In order to include the lowest and the highest cases that could *possibly* occur in a distribution, this distance must be laid off 5 times on each side of the mean. The scores can then be expressed in plus and minus values from 1 to 5, or they can be stated in plus values

with a zero point at the very bottom of the curve. In the latter case, the values are usually on a scale of 100 in order to avoid decimals. When plus and minus signs and decimals have thus been avoided, a convenient scale for research purposes is then available. This kind of score is known as a T score. A number of standardized tests give results in terms of T scores. While any form of standard scores has distinct advantages in terms of accuracy over other systems, there are disadvantages to be considered. For practical purposes, standard scores are very difficult to interpret. The very concept of the standard deviation cannot be explained in a few sentences. Only someone who has worked statistical problems with this concept rather extensively can give it much meaning. Teachers, supervisors in industry, government administrators, and the examinee himself profit little from an attempt to explain a T score of 40, as a rule. Most individuals who are interested in using test results, except research workers, do not like to become lost in technical verbiage beyond their comprehension. They are reluctant to acquire the extensive background necessary to understand it, and they may lose respect for testing in general, because it does not make sense to them.

There is an art in making test results meaningful to the layman. Much skill is required in order to prevent sound interpretations from being misused. The popular IQ has its faults, but it is so thoroughly absorbed into our culture as a means of expressing intelligence level that a substitute is unlikely for some time. Taking it at its face value without regard for qualitative observations and other data is a very common mistake. Every effort must be made to make norms meaningful. Simple language must be used in explaining scores. An intelligence-test report to be used by a high school or college teacher in some field other than psychology should not say, "The diagnostic indications of the Wechsler-Bellevue subtest patterning point to an anxiety neurosis. Therefore the obtained ratings are regarded as unreliable." This statement may be entirely correct in regard to the individual in question, but the person who will make use of the results will probably be quite confused by the professional language. A more direct kind of explanation might read something like this: "This man performed rather unevenly on the intelligence test. He was unable to do his best on several parts of it, probably because of an emotional difficulty that has been causing him con-

siderable anxiety. Therefore the examiner feels that the score obtained here probably is not as high as he might have earned had this personality problem not been interfering with effective thinking."

Among the kinds of scores that are relatively easy to interpret, percentiles are the most commonly used today. The median, or exact middle of the distribution, is the 50th percentile (sometimes abbreviated "centile"). In an exactly symmetrical curve, this value would be the same as the mean. The 75th percentile is a point three-fourths of the way up from the bottom of the distribution, or the point below which 75 per cent of the cases fall. The 99th percentile is similarly a point above which only 1 per cent of the sample population scored. Though these values are usually worked out in steps of 5, it is possible with a large group to break the curve down further into 100 such values (technically known as percentiles). The examiner who establishes percentile norms can explain them in the above manner or even draw a rough picture of the curve and the location of a score on it. Visualizing often helps individuals who may be poor at remembering figures. A precaution may need to be taken, however, in presenting the interpretation of these standards. If a test has been tried out on a select group, such as employees in offices, the norms for this group might be significantly higher than would have been the case had the test been tried on the general population. If Joe Dokes took two aptitude tests in the mechanical field, one of which had percentiles based on a large group of engineering students entering colleges of engineering, and the other had percentiles for high school seniors, the two could not be compared. Joe might be the 50th percentile for the general population and only the 35th for engineering students in the same capacity or achievement.

The problem of selecting an appropriate group for setting up norms is a very complex one. Since it is a little out of line with the purpose of this volume, it will be considered only briefly here. The reader is referred to Wechsler ² for an excellent example of the solution of such a problem on a large scale. A number of other examples of considerable merit can be found in the manuals of tests in a variety of fields. Factors that affect the mean and variability of

² Wechsler, David, *The Measurement of Adult Intelligence*, 3d ed. Baltimore: The Williams & Wilkins Company, 1944, Chap. 8.

scores include age, sex, geographic area, cultural background, socioeconomic level, rural or urban environment, education, experience, and a host of other less tangible circumstances. Motivation of the subjects is highly important, as mentioned earlier, and even the accuracy of large-scale scoring operations must meet a reasonable standard, as it often does not.

If the individual who constructs the test is unable to obtain a very large group that is in any sense "typical," he may use deciles rather than percentiles. Deciles merely reduce the scale to 10 instead of 100 to express a rough position of a score in a smaller population. If the resulting curve from a group of limited size seems to approach a normal distribution fairly closely, this finding at least indicates that the test is suitable for that particular group. If the curve deviates to a marked degree from the normal, either the test may be unsatisfactory in some way or the group may be decidedly unusual in some respect. Norms from small samples should be used with caution.

The term *standardized test* has been loosely used to cover measuring instruments that have reached widely varying degrees of refinement through scientific procedures. Indeed, some of the material that bears this label is scarcely worth the paper it is printed on. Broadly this concept requires that four essential conditions be met. The first of these is standard conditions of administration, discussed in Chapters 2 and 6. Briefly, this covers the giving of instructions, which must be done in the same manner to everyone who takes the test, and the amount of assistance that is given when difficulties arise. A number of other variables enter in here, some of which may be difficult if not impossible to control, but a test must include as many specific instructions for keeping conditions uniform as appear practicable. If too much is left to the examiner's judgment, the conditions will not be uniform enough for the results to be interpreted with assurance.

The second factor that is essential to standardization is uniformity of scoring, also considered in Chapters 2 and 6. Not only should a satisfactory key be provided, but some evidence of objectivity of scoring and fairness of right answers through item analysis should be available.

The third condition that is essential to standardization is adequate norms, considered in this chapter. The adequacy of norms is a sub-

ject of extensive controversy. Anyone acquainted with the broad field of existing test material and the considerable amount of literature on it will agree that it is not easy to prove that a representative sample has been secured, either of the general population or of some specialized occupational or educational group. In clinical groupings, such as neurotics and psychotics in various diagnostic categories, the problem becomes even more difficult through differences in theories concerning these disorders. Only a careful study of the materials, manuals, research, and criticism of each important and widely used test will enable the research worker in test development to justify his standardization acceptably. A specific problem of this kind could be mentioned. Public personnel agencies in several Northern states furnished data on examinations used locally for building-maintenance and skilled-trades positions. These examinations, together with the accumulated data, were secured by the Louisiana Department of State Civil Service through the Test Exchange Service of the Civil Service Assembly, in Chicago. Results obtained in Louisiana indicated that, not only were the norms far above the mean for the Louisiana groups in the same occupations, but the tests were beyond the reading-comprehension level for the group as a whole, and had to be greatly simplified. The differences in select groups of any kind in different geographic areas make the setting up of a "typical" population quite speculative.

The fourth criterion of standardization covers research to establish reliability, validity, and level of difficulty. Published materials differ to a marked degree in the extent to which this step has been carried out. The methods used in these investigations have been touched upon only briefly in this chapter, since most teachers and test technicians in government and industry seldom if ever have an opportunity to go far with such projects. Those who do must make an extensive study of statistics first, or else their investigations may contain serious errors that will bring justified criticism upon their work.

Most teachers and personnel workers will be primarily interested in evaluating the extent of standardization that has been done on the tests that they use. The tests they construct for their own purposes will usually not reach an advanced stage of refinement until after years of use. Few of them will become in a true sense thoroughly

standardized. However, as construction and improvement proceed step by step, every attempt possible should be made to meet the four conditions just discussed, in order that the interpretation of individual scores may be made meaningful.

Research. If apples, tables, automobiles, and chickens are added together, what does their sum mean? Other than a collection of so many miscellaneous unrelated objects, the total could have no meaning. If oranges, apples, bananas, and lemons are added, the total is meaningful to the extent that all of these are fruit. They are related; they have characteristics in common that are quite important.

If scores on such varied tasks as repeating digits, picture arrangement, vocabulary, and block designs are added together, the total has meaning, because these tasks measure closely related abilities. There is something in common among them that makes one correlate highly with another. This something is what we call *general intelligence*. If several tasks are closely related, that is to say, if a person scoring high on one is likely to rate high on the others, this finding needs to be explained in some way. To explain such an observation, authorities in mental measurement assume the presence of a factor. They may not always be able to define this factor clearly, but even if the verbal expression of its meaning is rather vague, it is often very useful.

Factor analysis is one of the most recent methods of research in test development. Complex as it may seem, its function is to simplify and clarify mental measurement. The use of it, as stated before, is largely confined to persons with extensive background in higher mathematics, yet the understanding of some of its basic principles is practically indispensable in order to follow much of what is now going on in test development. Only a few of the most highly skilled mathematical experts may be the persons mainly responsible for the most important future developments in test research. Probably only a limited number of users of tests will have the background to appreciate what is done, since most of them will become lost in the mathematical formulas. An attempt will be made here to explain only the most elementary concepts of the method.

In the above illustration of highly correlated mental functions from Wechsler's scale, the common factor may be assumed to be general intelligence. This assumed *g* factor, as it is called, may be a

sort of mental alertness, not directly measurable, but present in the performance of a variety of mental tasks. The *g* factor, then, is only indirectly measurable through performance on varied kinds of material on which success depends to an important extent upon *g*. There are other measures in which *g* would be much less important, as shown by much lower intercorrelations. General factors may not always refer to intelligence, but may be any function common to all subtests or all members of a battery of tests. When general factors are prominent, total scores have meaning. When there is a very low correlation between any two parts of a battery, the sum of the scores on the parts becomes less applicable than the individual part scores. On Seashore's Measures of Musical Talent, for example, the general factor is insignificant, since sense of pitch, loudness, time, timbre, rhythm, and tonal memory are relatively independent of one another. No total score is used. The shape of the profile forms the basis of interpretation, not a single total rating. Thus a general factor may be present, but weak. It may be easily identified or too vague to be expressed adequately in words.

A second kind of factor is found in some, but not all, members of a battery. This kind is called a *group factor*. It may be present in two, three, or more parts of the battery to a strong or weak extent. Sometimes two parts correlate so highly that they are measuring the same trait. Then one or the other can be eliminated without impairing the validity of the battery.

A third kind of factor is found only in a single subtest or member of a battery. It is termed a *unique factor*. The Seashore Measures of Musical Talent have important unique factors. Each part is relatively independent of every other part.

A battery may contain only one of these classes of factors. Only unique factors may be present. Another battery may measure the general factor alone. Still another battery may contain unique and group factors, and it is possible to have all three types represented. A variety of combinations is possible. The relative importance of any one kind may be found by statistical methods. Some tests are relatively pure or saturated with one factor. Others measure several factors, some of which are more important than others. In the United States Air Force, much research is in progress with the aim of identifying, clarifying, and defining these factors. For a full dis-

cussion of the techniques being used in such research, the reader is referred to Thurstone.³ Thorough acquaintance with his methods will enable one to read with interest the numerous studies in psychological journals that deal with the application of factor analysis to various tests.

This abbreviated volume has not covered comprehensively the statistical tools employed in the later stages of test development. Such tools have been fully described and applied by many writers in excellent basic texts dealing with these problems. This volume has attempted to explore fully the earlier stages entirely neglected or insufficiently covered elsewhere, since the users of this work will in most instances be primarily concerned with these. In summary, the steps covered may be stated as follows:

1. Planning the examination as a whole so that it will be valid for its purpose, including outlining on the basis of course content or job description
2. Shaping up each item in such a way that the task will be clear and the scoring objective
3. Providing as far as possible for uniform conditions of administration and reliable scoring
4. Obtaining adequate review and expert assistance where needed
5. Tryout on a suitable group
6. Item analysis that will result in improvement when necessary
7. Setting up useful tentative norms

If these seven stages are thoroughly and properly executed, later studies of validity and reliability together with factor analysis will be likely to produce more gratifying results than would be obtained if actual construction were carelessly and unskillfully done.

³ Thurstone, L. L., *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947.

SAMPLE PROBLEM IN TEST CONSTRUCTION AND SOLUTION

IN AN effort to bring theoretical principles down to earth, the writer has added here a project more complete than the fragmentary illustrations in various chapters. This outline of a specific project and the stages of its solution as far as the writer had an opportunity to carry it is not intended as a complete research study. The discontinuance of the Louisiana Department of State Civil Service at an inopportune time (autumn, 1948) prevented conclusion of studies of reliability and validity then only started. In the literature on industrial and educational testing programs may be found more conclusive evidence of the ultimate value of the measuring instruments used. However, this project in test development was a thorough preliminary. The writer wishes to acknowledge the valuable assistance rendered in planning this project by other examiners then in the agency, particularly his supervisors at different stages of it—Donnell Read and Norman Ecklund.

The wartime shortage of civilian manpower at the time made it essential that state employment be made attractive, including, of course, examinations, without, however, lowering standards too much so far as permanent employment was concerned. Therefore, face validity (the appearance of relatedness to the work) was an important consideration.

With this background in mind, the writer planned a selection procedure for clerks, typists, and stenographers at three levels of responsibility. For typing and stenographic positions, performance tests were included, but the details of these will not be discussed here. Written tests were identical for clerks, typists, and stenographers at each of the three levels. Entrance levels for all of these three classes

were given the same written material, etc. The second level overlapped the first, but included more material, some of which was at a higher level of difficulty. The third level overlapped the other two to an extent, but again added more complex items in somewhat different subject-matter areas.

The part of the procedure to be taken up in detail here consists of the examination for one class, Clerk III. Jobs in this class were either supervisory or highly responsible office work, nearly always involving some public contact as well as working with subordinates. These employees supervised filing, typing, office-machine operation, and other routine clerical work. Examples of duties included writing reports, simple interpretation or application of state laws, dictating correspondence, checking work of subordinates or others for accuracy, and meeting the public. Not all employees in this class had responsibility for every one of these duties, but usually most of them were a part of the job. Some knowledge of office procedures, correct business English, and general vocabulary was considered essential. Tact and courtesy in dealing with people, ability to do clerical work rapidly and accurately, and good judgment were required. These were expected at the time of appointment. Presumably the laws, rules, and procedures pertaining to the particular department could be learned on the job. Three years of office experience, one of which was in a supervisory or other responsible capacity, was the stated requirement. High school graduation was considered desirable, but not required.

With this review of the specification in mind, the writer supplemented the observations made by the classification technician by observing some of the work on various jobs in this class himself. In this way he learned what characteristics were common to at least a sample of the jobs. The essentials in the specification seemed to be a sound basis for planning the examination. A tentative outline was then drawn up as shown on page 181.

Although the personality evaluation and rating of training and experience from the applications presented many problems, the details of these would be outside the scope of the present volume. Therefore, discussion of them beyond the preliminary outline stage will be omitted except for the statement that each of these parts had to be passed by an applicant in order to be eligible for employment.

	<i>No. of items</i>	<i>Form</i>
I. Written test		
1. Clerical aptitude, names and numbers	150	Matching
2. Vocabulary, slanted toward office work	20	M.C.
3. Following directions	15	Matching
4. English usage	15	M.C.
5. Office procedures, supervision, etc.	10	M.C.
(Clerical aptitude 10-minute speed test. Remainder 2-hour power test.)		
Weight of entire written part	40 per cent	
II. Rating of training and experience		
This was to be set up by dividing experience roughly into three categories, A, B, and C, of decreasing value as to quality, allowing greater credit for more recent experience than for less recent experience within each category. Small bonus credits were to be allowed for each year of college education and for completion of high school		
Weight	30 per cent	
III. Oral examination or interview		
This was to be of 15 to 20 minutes' length, conducted by a board of 3 members, and designed to evaluate personality and attitudes toward the work. Ratings to be made independently by each member on forms, then averaged. Traits to be rated were:		
1. General appearance, grooming, clothes, etc.		
2. Voice quality		
3. Tact and courtesy		
4. Mental alertness		
5. Facility in verbal expression		
6. Poise and self-assurance		
7. Potential interest in the work		
8. Emotional stability		
This was to be supplemented by a total estimate of fitness for the job, which, like the 8 traits, was to be marked on a 5-point scale.		
Weight	30 per cent	

Of course, the minimum qualifications for entrance to the examination gave him 70 per cent on experience. Anything above that minimum raised the score. In the case of the interview, the requirement that it be passed with a rating of 70 per cent or better was essential. There are intellectually brilliant individuals who might be admitted and might score almost 100 per cent on the written part, yet be totally unfit as supervisors because of emotional problems of a severe nature. Such a person could fail the interview and still make a rating above 70 on the examination as a whole were this provision not made in the announcement. Such cases have made quite notorious failures on the job, and have antagonized everyone or destroyed good morale. The writer would not, on the basis of his experience, guarantee that

such a maladjusted individual could not force himself to behave well and fool the interviewers. This precaution makes the employment of such a person very unlikely, however.

The aptitude and achievement parts of the selection procedure are of most concern here. These will be discussed step by step. For the measurement of clerical aptitude, the writer decided upon the use of a modified form of the name- and number-matching method used in the Minnesota Test for Clerical Workers. This modification provided for transferring answers to a machine-scored answer sheet. Later the examiner was able to demonstrate that his modified, machine-scored form correlated .94 with the original Minnesota test. For large-scale scoring purposes, it had a decided advantage. Material for the names was picked at random from a telephone directory, while the numbers were written on a random basis, with series of different lengths. Practice exercises for both names and numbers were given on the cover page. Ten minutes was allowed for working columns of names and columns of numbers alternated, thus giving only one score and one timing period.

Materials for the vocabulary section followed a multiple-choice form, and were secured from a variety of sources. Other civil service examinations secured through the Test Exchange Service of the Civil Service Assembly were of some help, but the best source was, perhaps, words commonly misused or used in the wrong context in business correspondence. Such words were observed by several instructors in business-letter writing and by the examiner himself. Thus the vocabulary section was not too academic. It was developed not from the dictionary but from the practical setting of the office, to a large extent.

The following directions and English-usage sections were planned as described in Chapter 4, the latter again being drawn from the context of business correspondence. In the office-procedures and supervision items, the greatest amount of copying from other tests occurred, but neither the copied items nor those original with the examiner were entirely satisfactory at first, particularly in the field of supervision. Many modifications of the material in this last section were tried.

Weighting of different sections according to their relative importance was considered, but this would have introduced complications into scoring which would have resulted in loss of speed and accuracy in large-scale operations. Therefore, each section was ultimately allowed to weight itself automatically by the number of items in it. Although different numbers of items were tried in each section, the office-practices and supervision section was shortened so that it would have less weight in the total score. This was not done until a preliminary tryout showed this section to be inferior in that it did not discriminate so well as the others. It was retained for its face-validity value chiefly. Some arithmetic problems might well have had more actual validity, but they were tried at one time and found to be rather unacceptable to candidates. The preliminary weighting of the sections therefore was only subjective and tentative.

Instructions and practice exercises were constructed next. These had to be clear and simple without consuming too much time. An attempt was made to anticipate in advance the questions that would be likely to arise and to answer them in the instructions. The final product, to be read aloud while candidates followed silently, was complete and even repetitious on important points. Superior candidates would doubtless experience considerable boredom, but mediocre and inferior ones, if they were attentive, could hardly ever fail to understand how to record answers. Thus the entire examination was felt to be fair to all who met minimum qualifications, since it was relatively free from verbal monstrosities that might be considered too academic.

The tentative outline for this examination worked out quite well, since an abundance of material could be found for each section. This is often not true, however. On some examinations the author has found such a scarcity of material that could be made into good items on some particular topic that the topic itself was abandoned from the final outline. In this case the topical breakdown remained unchanged. Only the length of two sections changed from the time of planning to the stage of completion of the rough draft of the examination itself. The vocabulary portion was shortened and the English usage

lengthened. There were some problems as to what to include that could not be solved before the test was tried except by subjective judgment of the writer and consultants as to their merits. If radical disagreement turned up among the reviewers, the item in question was omitted.

Two consultants were secured who had accumulated a great deal of experience in office supervision and who were both considered exceptionally efficient in their own work and effective in supervision. The opinions of several members of the staff of the personnel agency were essentially in agreement upon the merits of these two. Their review was oriented toward an appraisal of the material from the applicant's angle. Both expressed interest in giving this assistance when interviewed by the examiner. Neither was content to label an item as "satisfactory," but reasons for inclusion or exclusion of each question were often clearly stated. Sometimes one of the reviewers could not state exactly why he did not like a question included, but usually the suggestions given were clear and constructive. An instructor in business-letter writing worked closely with the examiner on the English sentences. Each reviewer occasionally wrote a substitute item after giving an objection to one submitted to him for consideration. The substitute was often in need of revision as to form, but it usually brought in a new idea which probably contributed a fairer sample of the essentials of the job. As might be expected, the language used by the examiner was frequently too academic, even though his basic ideas were almost always judged to be quite practical. The examiner's background of training and experience could not be escaped altogether, and the reviewers were most helpful in finding defects of this sort, which would probably never have been evident to the writer, until, perhaps, after an item analysis had been made.

After the reviewers' suggestions had been followed to the best of the examiner's ability, the examiner himself found that there were a few technical and typographical errors. One consisted of a pattern of answers on the key, apparently unnoticed before, that went 2, 3, 1, 4, 2, 3, 1, 4, etc., for some 15 items. Rearrangement of choices in a few spots avoided the possible clues to correct answers. Other examiners were asked to review the material from a construction

viewpoint, and two of them did this in detail. Several valuable suggestions resulted.

After duplication of the final forms, the material was used on a state-wide basis for the first time. The number of applicants was not large enough for statistical item analysis, but inspection of results led to a few minor changes. The examination was advertised several times, and enough data were assembled to make an item analysis possible, but only a part of the material was ever handled in this way. Several items considered good by the examiner and two reviewers did not yield good results, particularly in the office-procedures and supervision material. The distributions of scores approached the normal curve, however, as closely as could be expected from a total of approximately 400 cases. Therefore, the test was probably the right level in difficulty and of some value in discriminating among the applicants those of various levels of probable fitness for the jobs.

Unfortunately, the passing point could be set only arbitrarily on a statistical basis at one sigma below the mean. This point was regarded by the two reviewers as reasonable after examining papers barely above and barely below this level.

Figures on the reliability of the test as a whole were not computed at the time the agency went out of existence, though the results on the clerical aptitude section (split-half) were in the low 90's on two trials on two different groups. Validity could not be found until a rating procedure could be set up, and this final stage was in progress when the writer resigned his position as examiner.

The partially completed project presented here stresses the careful work on the earlier steps in the process, apparently often passed over lightly in the literature on tests as if it were unimportant. By their very nature, most teaching positions do not allow time for thorough test development to the refined stage of standardization, and indeed few personnel technicians regularly have an opportunity to carry through as far as desirable. However, time spent in preparing good tests is time well spent. General respect for and cooperation in taking tests will be achieved only if people are convinced that tests are good enough to be worth taking.

SAMPLE ITEMS

1. *Clerical aptitude, numbers:*

INSTRUCTIONS: Place a check mark (✓) between the pairs that are exactly alike. Indicate *number* of check marks on answer sheet.

26498 ✓ 26498
 386 _____ 366
 9185467 _____ 9185367
 625489 ✓ 625489
 377 ✓ 377

Answer Sheet				
1	2	3	4	5
⋮	⋮	⋮	⋮	⋮

2. *Clerical aptitude, names:*

D. F. Baldwin Co. ✓ D. F. Baldwin Co.
 Peter Mac Donald _____ Peter Mc Donald
 Goodwin Printing Shop _____ Goodwyn Printing Shop
 Milton Osborn ✓ Milton Osborn
 Anderson Bros. _____ Anderson Brothers

1	2	3	4	5
⋮	⋮	⋮	⋮	⋮

3. *Vocabulary:*

Verify 1. record 2. postpone 3. calculate 4. prove

1	2	3	4	5
⋮	⋮	⋮	⋮	⋮

4. *English usage:*

INSTRUCTIONS: On the answer sheet, indicate the *number* of the section containing an error. If the sentence is entirely correct, mark 5 on the answer sheet.

1 2 3
 The man whom/ we believe will/ accept appointment to this position/
 4
 has had considerable experience.

1	2	3	4	5
⋮	⋮	⋮	⋮	⋮

5. *Office practices:*

If a typist is constantly making the same kind of error, ordinarily the office supervisor should *first*

1. reprimand the typist in front of the others in the office
2. give her written notice that if she does not improve her work, she will lose her position
3. report the matter to a higher executive
4. call the typist's attention to specific examples of her errors.

1	2	3	4	5
::	::	::	::	::

EXAMPLE OF COMPARATIVE ANSWERS FOR A PERFORMANCE-TEST PROBLEM¹

Organization of Materials. On the following pages are given the statements of one client and those of nine student counselors. The counselors' remarks are those of nine advanced students in a graduate course in counseling in the Department of Psychology of Baylor University. The client's statements are those which occurred in an actual interview, as originally conducted. Then follows each of the responses of each of the nine student counselors. The numbers assigned to the students have no significance. The sequence is random, and covers a wide range as to quality of counseling. The code of each individual student counselor's remarks, indicated by sc1, sc2, etc., merely serves to identify the particular student counselor in each of 20 response situations.

Instructions to Examinees. The instructions were in general (1) that each student would perform the make-believe counseling as a part of his final examination in the course, (2) that each would be independently rated, relative to the others, by three professional psychologists, (3) that in each case the identical "client" statements would be used, (4) that the client's statements would be those of a real client in a real interview previously recorded and edited slightly for English and wordage, and (5) that the make-believe examiner would, of necessity, be not allowed to deviate from the actual client's statements. The interviews would, therefore, appear to be slightly disjointed when recorded and transcribed for the individual student

¹ Material prepared by H. L. Trites under the supervision of Carl Rogers.

counselor. Having been thus informed, the examiner read the following statement to each student counselor:

In this counseling interview, the case has been referred to you by a co-worker. He (the client) is a Negro boy in an educational situation. He seems to have personal problems that are blocking his satisfactory progress.

I will read the client's part, and you are the counselor. The remarks of the client, which are read to you, will form the basis upon which you are to conduct your counseling. The approach to be used is nondirective, and you will be graded on the degree to which you conform to this school of thought in handling the various problems and situations encountered in this interview.

The last sentence of the instructions would be essential in a personnel-selection situation. It was omitted in the original use of the material quoted here, because the class had already been oriented to the nondirective approach throughout the course.

Subsequent Treatment of the Test Materials. While the make-believe interview was in progress, the entire conversation was recorded. It was next transcribed literally. Then it was edited somewhat, but only slightly so far as the statements of the examinees were concerned. These "responses" were separated from the common statements of the client (read from the original interview by the examiner). The student counselors' responses were finally arranged by client statements into 20 responses to 20 client statements. In Appendix C are reproduced three complete interviews as they occurred in the performance test itself.

Suggestions for Study. The materials may be read straight through, as presented, to compare the various responses of the nine different student counselors to each statement that was made by the client.

The comparative quality of each candidate counselor may best be appraised by reading collectively the responses of each individual through the entire interview (see Appendix C). That is, by putting together the client's remarks and those of student counselor No. 4, a complete transcript of the entire interview can be constructed, as it would have occurred if No. 4 had been the counselor. Likewise, a complete interview with No. 7 as counselor can be obtained by taking the client's statements and No. 7's responses.

Case of Mr. ——— with various responses.

CLIENT. I would like to be a psychologist. I understand you're a psychologist. I'm working in the daytime, I'm going to school at night. And right now it looks like that's what my whole life is wrapped up in education. That is the sole thing I want. And I want to find out whether I'm capable, whether I'm mentally capable. I seem to be, but I'm not sure. I have a doubt, and Mrs. ———, she was my general science, general psychology, yes, my general psychology, instructor, she gave us some type of IQ. I don't know exactly what type it was, but it did prove very favorable. But she said it was just an experiment, nothing for me to take definitely. Well, I would like to have some kind of definite understanding there of my own capabilities. I did very well in high school and I seem to be doing very well in college, but with all the money I'm going to have to spend and the hopes I'm putting in my future through education, I would like to know just where I stand.

Responses:

- sc1. First of all, I'd like you to know that I'm a counselor, and it is my job to listen to your troubles and try to find some methods or means to help you solve them. Now these doubts or fears that you have in your inner nature are things that we can air, and possibly help out in our discussion. Could you possibly tell me more about yourself?
- sc2. No, I may inform you that I am not a psychologist. As you have stated that you would like some sort of testing to see just where you stand. There are many different types of tests that perhaps we may give you—before we go much further in this interview.
- sc3. In other words, you have fairly definitely chosen your objective and you are very interested in it, but you are not sure what your capabilities are and you want to find out for sure before you expend all this effort. Is that right?
- sc4. (I didn't catch anything on that then.)
- sc5. You really want to find out what you—uh—what might be best for you. Is that right?
- sc6. You're really concerned about whether or not you have the ability to take this education. Is that right?
- sc7. Then you feel that you would like to summarize, or get together all your thoughts on this so that you can determine whether or not you have the ability to make a good psychologist?
- sc8. You're a little unsure as to exactly just what you want to do. Is that right?

sc9. You believe your interest in psychology that you have come in contact with recently is something that you want to go on with?

CLIENT. The kind or type of work I want to go into is child guidance work. From what I understand, all this is just through what I have asked questions about and so forth. I have no facts or anything. But I talked to Father _____ about it and around the Center and the YMCA, etc. There are not very many Negroes in that specific type of work—child guidance—and I do have that handicap to come over, to get into the field to break down the barriers. I don't mind all of that if I know that I can get somewhere. I don't want—I want a livelihood but I'm not in it to have a big house and a lot of money and everything. I want to go into it because I think I like it.

Responses:

sc1. You feel then that your main idea in going into the work is to satisfy your inner nature and not just your economic status.

sc2. You lack a particular background for this type of work, outside of the fact that you feel you would like to go into it for a livelihood.

sc3. As I understand it, then, the main thing you want in this occupation is job satisfaction.

sc4. Well, do you feel that you can make a go in this field?

sc5. You feel that you would like child psychology?

sc6. You've given this problem a lot of thought. Would you like to—uh—express your opinion about it a little bit more?

sc7. Then you feel that the fact that you are a Negro would hinder you in your work as a psychologist. Is that right?

sc8. You're really interested in that kind of work.

sc9. Child guidance is one field in psychology that interests you more than any other.

CLIENT. That's right. I want to be confident in myself and it doesn't matter what I have to face.

Responses:

sc1. In other words, you have indecisions now and then, about your coming work.

sc2. (*Pause—long pause.*)

sc3. You want to be sure you are doing the right thing, then, before you start.

sc4. (*Pause.*)

sc5. You're not so sure.

sc6. You feel that it will be a success and it is worth it.

sc7. Is it just the problem of the lack of self-confidence? Is that your problem?

sc8. Yes? (*Weakly.*)

sc9. You feel that you are willing to go into all that to find out what you can do for yourself and others too?

CLIENT. That's right. Because I mean it's not . . . it's a sacrifice now. I work in a factory, and, well, with working and studying until two o'clock in the morning, and my physical condition is good, but I am extremely nervous and it does slightly affect my heart and I did have ulcers when I was eighteen and . . . I never had operation, but I had a milk diet and all that which never really cured them up. All of that, my mental situation and my physical situation, my nerves and everything, I want to know that I can do it on my own regardless of the obstacles if I have confidence in myself. So I'd like to have some type of IQ, if you can give it. I don't have any money, but whatever the charge is, I'll try to meet it. Miss _____ at the time told me they were very expensive, but she gave me this one free. Took about five hours one evening, and she said it was very good. But that was all. She didn't tell me just where I was weak or what I was good in and so forth.

Responses:

sc1. You feel, then, that the main thing you are trying to strive for is confidence in yourself.

sc2. First, I think that we should give you some more tests, to see what particular thing that you are best fitted for. You may not even—you may be fitted for something else outside the field of child guidance that you seem to be so interested in.

sc3. Then, actually you feel that there are a number of factors that enter into going into this particular occupation—your health, your mentality, and all these things which make up—uh—what you want to do in this matter.

sc4. What do you feel this nervousness is caused by?

sc5. I see-e-e. (*Deliberately.*)

sc6. You're in doubt as to just which way to turn now.

sc7. From your conversation it seems as though you are kind of mixed up as to whether or not you have the physical ability to do this job. It seems that your physical ability wouldn't need to be considered here as much as the testing. Now, is it that you want to find out from

the test whether you are going to be able to be a psychologist, or whether or not you would make a good one?

sc8. You're—you like to know just where you stand.

sc9. You realize, as I get from what you said, that there *will* be problems to be faced.

CLIENT. M-hm. Well, I wonder if we had the results of the test that Miss _____ gave me, if we had them available during the counseling, could we derive benefit from them? Could we use them . . . could they possibly do a lot to clear up some doubts?

Responses:

sc1. As I see it, these doubts and worries that you have in your mind disturb you a great deal.

sc2. What makes you particularly think that you would be good in this type of work even though—that there are only a few colored people that follow the child guidance program?

sc3. I see. You're leaning rather heavily on this IQ. You feel like it is one important factor in arriving at this decision . . . if that information was available.

sc4. That is possible.

sc5. You're just not for sure whether you are physically able to move into this type of a training program.

sc6. They might, but it seems to me that the most important thing would be what you actually believe about yourself.

sc7. Well, they perhaps could. We'll just have to see about that. Would you like for me to give you a test to see what your mental ability is?

sc8. They may. We'll look that up later.

sc9. In other words you would like to find out more of your capabilities.

CLIENT. Yes, helpful to me, but along with what you are doing. Because I don't know, I mean maybe, maybe, but to me the whole thing is tied up together, all this, the emotional and the intellectual may be that together, maybe not, I don't know. But you can appreciate my position when you realize that on my education, I feel like that's my whole future. I don't definitely, I definitely do not want to work in a factory all my life. I mean, I believe I could go insane doing that. And there seems to be nothing else in life for me but that. I mean I do want a home and a family and so forth, but to have a home and family I've got to have something else too. And it seems, everything now seems to be education, but there seems to be some emotional or psychological obstacle or barrier or frustration. And I was

wondering if somewhere that is tied in also. That's not the main thing, that's just something I was wondering, you see, and, and if possible could we use the IQ test to possibly open up a road to what really *is* bothering me emotionally.

Responses:

- sc1. As I see it then, these frustrations and inner feelings that you have, in your own mind, are the deepest things concerning you at this time.
- sc2. That depends upon the type of test that she gave you and I would not be stating that they would absolutely give us the correct answer.
- sc3. Well, we can try this IQ test if you want to. However, as you state there, there are emotional factors and other things that you are interested in exploring too, and you might cover some of those things in this session.
- sc4. What do you feel this emotional difficulty is caused by, chiefly?
- sc5. Do you feel that the test you have had might be of some value to you in this situation?
- sc6. You feel you want the better things in life, yet there is something in the way and you don't know just what it is. Is that it?
- sc7. It seems that what you want is to be sure of your security in life. That could possibly be your trouble. Is that it?
- sc8. You're wondering if the IQ tests will give an indication of your emotional problems.
- sc9. This information probably would give you some idea as to whether you thought you would be able to do this work or not.

CLIENT. That's right. That's what I told Miss _____ when I talked to her.

Responses:

- sc1. I see.
- sc2. It may be of some help to us in this particular solution. I wouldn't say altogether.
- sc3. It worries you quite a lot then. You feel that they may indicate some emotional difficulty that you can't analyze.
- sc4. I see.
- sc5. You feel like the results of the intelligence tests you have had might be of some help to you in reaching a definite conclusion.

- sc6. Would you care to tell me a little bit more about your problem?
What seems to be the trouble?
- sc7. Would you care to tell me a little more about what you and Miss _____ discussed?
- sc8. Yes. (*Weakly.*)
- sc9. Perhaps you thought you would be able to be more sure of yourself if it was possible to take those tests.

CLIENT. That's it. The whole thing, I am rather convinced it's an emotional setup, because I'm nervous, and I dream at night, I'm exhausted and I really sleep; I mean, I'm very sleepy all the time though I try to get eight to nine hours' sleep at night. Some nights I can't because of school, but other nights then I'll get maybe ten or eleven hours, something like that. But somehow I try to get enough sleep. But I dream constantly. Crazy dreams that I know must have as a . . . as I said, I'm going to major in psychology some day, so I'm . . . I do a little bit of reading in psychology. Possibly it is wrong, possibly because of my lack of knowledge I'm drawing the wrong conclusions, but I believe these dreams are the results of some emotional stress.

Responses:

- sc1. This emotional stress that you think about, and these dreams and everything are pretty well tied together in your emotions. Is that how you feel?
- sc2. Well, now just tell me, on this test that Miss _____ gave to you—what did she tell you about this test? Give me a little more of it.
- sc3. This experience with these dreams has left you a bit confused about just where you stand. Is that it?
- sc4. You can't quite figure out what's behind all this emotional stress, can you?
- sc5. I see.
- sc6. You seem quite concerned about your emotional and mental condition.
- sc7. Then you feel that you are emotionally upset. Is that it? You feel that because you are emotionally upset, it might hinder your progress as a psychologist.
- sc8. You're interested then in the connection between your dreams and psychology as you like to study it?
- sc9. Evidently you enjoyed taking those tests and taking the same type probably would not bother you in the least. In fact from what you say it seems as though it is the thing you want to do.

CLIENT. As we talk, I'll be able to bring them out more clearly, but usually when I'm in an emotional upheaval, I have more of these dreams. Just a period about, oh, about three weeks ago, up to that period for about six weeks, well, I had final exams at school and everything was going along placidly, though it was final exams at school and everything and I made B's and A's in my grades and so forth. But then all of a sudden a disturbance came in my life and it was presaged by a period of dreaming, even before I knew what was going to happen. That made me sensitive to some type of trouble. And the greatest thing that happened in my life happened to me, detrimentally, and, well, all that, it just doesn't tie up or it does tie up, I don't know what.

Responses:

- sc1. You feel, then, that there is a definite connection between your dreams and your frustrations and also your ambitions.
- sc2. Perhaps we should refer you to one of the psychiatrists here in the building. If you have a few minutes, I'll try to make an appointment for you.
- sc3. Well, actually, I'm no Mr. Anthony, to just hand you out an answer here but I think in doing just the thing you've been doing . . . in exploring these attitudes and these feeling that you can arrive at a decision that would satisfy you better than any I could hand out.
- sc4. Could you elaborate a bit more on this disturbance you had?
- sc5. You feel that the dreams you are having might have some connection with the emotional upset that you are worrying about. Is that the way you are feeling about it?
- sc6. You feel that, perhaps, these dreams have something to do with your trouble?
- sc7. You seem to think that there is a definite relationship between your dreaming and the experiences that you dread. Now would you care to elaborate on that a little further and tell me more about your dream so that I can analyze this a little better.
- sc8. You believe then there is some connection between your dreams and your problems as they occur?
- sc9. You certainly show that you have been doing some independent thinking, and you want to arrive at your own conclusions as to the solution to your problem.

CLIENT. Well, now, I mean, I'm about talked out. (*Laughs.*) I mean I could talk more but that was by way of introduction, so that you could find out just what I want.

Responses:

- sc1. It's more for you to say exactly what you want. More for me to sit here and listen and try to reflect back to you what we are trying to do here.
- sc2. From what you tell me it seems as that there is something else here that seems to be bothering you outside of this particular degree or the work that you want to get in child guidance. Now we may be able to get down here to a final answer if you just come out and put everything on the table.
- sc3. (*Long pause.*)
- sc4. You don't feel you would like to discuss it further?
- sc5. You're not so sure whether there is any connection between this disturbance and your dreams.
- sc6. I see.
- sc7. Well, would you like to come in at another time and discuss this with me further?
- sc8. H-um, I see.
- sc9. Of course, if you continue to give it your best thinking, some good will come out of it.

CLIENT. That's right.

Responses:

- sc1. Yes, I believe that if you gain a deeper insight into your own problems and understand your own ways of thinking better you'd be in a much better position to understand the difficulties you are going through.
- sc2. Just a minute, I'll call Dr. X and make an appointment for you and you can go on down and see what he has to talk to you about.
- sc3. (*Pause.*)
- sc4. (*Pause.*)
- sc5. Well, perhaps if we can go into this just a little bit further, perhaps we can reach some definite conclusion.
- sc6. (*Pause.*) You feel that these dreams are really having an effect on your emotional life.
- sc7. How about coming in at your next free hour, and we'll go into this a little further; and in the meantime you'll have some time to be thinking it over again and perhaps you can elaborate on it a little further at the next session.
- sc8. (*Paused and blocked.*)
- sc9. Well, I can definitely see that you, as I said a while ago, you have an

idea that you do have a problem, and that the good thing in your favor is that you want to find a solution. Of course your coming here to talk about it and discuss the thing you have already done will put us in a position to assist you as much as we can.

CLIENT. That's right. That's right.

Responses:

- sc1. Now, can you tell me a bit more about these disturbances you spoke about before?
- sc2. He—uh—he's out at present, right at this particular moment and the girl said it would be all right for you to come right on down there.
- sc3. (*Pause.*)
- sc4. (*Pause.*)
- sc5. Well, tell me, tell me just a little bit more about this situation.
- sc6. Would you care to explain any further?
- sc7. Uh—when would you like to come in?
- sc8. (*Paused and blocked.*)
- sc9. Have you made any definite plans as to when you would like to come back to talk about this thing? (*Pause.*) Some more?

CLIENT. I don't know just what causes it. I mean, it may be something in my personal life, or it may be something that happened to me that I have tried to talk in my mind and bury away, and subconsciously I keep holding it down, but it wants to come to the front. And I don't know what it is, and I wish I could find what it is and solve it once and for all, and everything would dissolve. I don't know whether it's financial or, I mean, I don't know what it is. But it's something. I mean it's *got* to be something.

Responses:

- sc1. In other words, once you find out that which is bothering you, you'll be in a better position to understand your own inner nature. Is that what you mean?
- sc2. Yes, there must be an answer to all questions. And it is now about time that the doctor should be back in his office, so we'll just go right on down.
- sc3. There's something deep-seated in your personality that you feel if you could bring it out in the open, you could whip it, and solve your problem. Is that it?
- sc4. This very probably can be tied in with the emotional disturbance you mentioned earlier in the conversation.

- sc5. Even though you try not to think about it, it still develops sooner or later.
- sc6. You feel you have a problem somewhere, but you don't know what it is, yet you'd like to find it.
- sc7. Then you feel that there is something that you can't put your fingers on that is bothering you. Is that right?
- sc8. You feel that your problem is almost an obvious one.
- sc9. You're right, everything has a cause. And working at it as you are, I think you'll come to a conclusion as to what is the best thing to do.

CLIENT. That's right. I have a friend (*pause*), well, anyway, he and I do a lot of talking. And there's something that we have a mutual agreement on. I just don't know what it is. But I just met him about three years ago. And our friendship has increased. We enjoy sitting down and talking to each other. He has the same type of problem. We just don't know; we tried to sit down and just figure out what it was. He wants an education and he's not getting anywhere. I want an education. It looks like I'm getting the education, but I don't seem to be getting anywhere, even though I'm getting an education. Maybe that's just a belief of my own. Possible. But I'm twenty-seven and all I have is a car and working for an education.

Responses:

- sc1. You feel, then, in your opinion in your education you'll obtain everything you'll need toward the goal you are attempting to seek.
- sc2. Well, after you have had a chance to talk to Dr. X, perhaps he'll be able to enlighten you on some of these personal problems that we haven't gone into, that seem to be bothering you.
- sc3. You've been trying to do something about this situation for some time then, you've talked it over with your friend and you've tried through education to solve it, but so far you haven't been entirely successful.
- sc4. Are you real close to this friend you are talking about?
- sc5. You seem to be headed in some special direction but you just don't know where you're going.
- sc6. Sometimes it all seems futile to you. Is that it?
- sc7. Then you feel that your lack of progress—uh—is due to the fact that there is something that you're not getting out of your education. Is that it?
- sc8. You're interested then, in talking with your friend in trying to solve your problems.
- sc9. Well, of course, as you go on through life you'll better your chances.

CLIENT. That's right. And I believe that maybe it's because I doubt whether this is the path . . . I believe it is . . . doubt whether this is the path or whether I'm capable of doing what I want to do when I am going through this way or what, I don't know. That may be the doubt or as I say, it may be something else in my life. Really, I don't know what it is. And if maybe just a big (*pause*) but I don't think so, because it affects me too much. I make friends and I'm always at odds when I don't see things their way, so consequently, I mean my friendships are temporary. And everybody tells me I'm radical, I'm, I'm idealistic all that. And I don't know whether this one fellow and I remained friends because we have a common problem or a common understanding.

Responses:

- sc1. Could it possibly be that the distinction you make between your friend and your problems are tied up with your over-all problem which you are attempting to solve? Could that be the situation?
- sc2. There seems to be some doubt in your mind, about this particular work that you spoke about . . . this child guidance . . . and whether you would be able to go on with the work if you did have the proper qualifications.
- sc3. This undefined thing in your personality then has—uh—led to difficulties so far as adjusting with people and friends you make, and all. Is that right?
- sc4. You find it very difficult to make friends with anyone except this one person. Is that right?
- sc5. You have a feeling that your friend understands the way you feel about these things.
- sc6. You feel that if you could bring this problem to the front and solve it, it would help you. Is that it?
- sc7. Then you doubt—as if—you doubt whether or not the friendship is a true friendship. You're disturbed about your—the—emotional setup here. Is that it?
- sc8. You're interested in keeping up your friends that you meet and you do make friends easily but this friend of yours still seems to be the one you like to talk to and understand the most.
- sc9. Having a friend that you are able to talk things to is really quite a consolation and I'm sure you're very happy that you do have one.

CLIENT. Well, I believe it *is* education. I mean, I don't see right now anything else I want. What type of education . . . Am I in the right education? I know it's not science, because I'm no good at math, or anything

like that. In philosophy in school I excel. In psychology and sociology too. So it must be something along the human interest field.

Responses:

- sc1. I see, well, could you tell me a little bit more about your interests?
- sc2. After you have taken these tests that we are going to give you, the final results will give us something to work on as to whether you are qualified to go into this type of work that you seem to be so interested in.
- sc3. You feel, definitely, then, that it is along the line you outlined to me here, but you are still not too sure where you should go from here.
- sc4. I see.
- sc5. You have a feeling that your education is absolutely necessary to give you the desires that you have in life.
- sc6. Um—hm.
- sc7. Then you feel that perhaps that the education you are getting is not the type that you need to get. Yet you cannot see why the type of education you are getting is not the type that you need. Is that right?
- sc8. You are, then, mostly interested in people as you meet them.
- sc9. To succeed in the field you definitely have to have interests in other people because you are going to be called on to assist them in finding the solution to their problems.

CLIENT. That's right. Like I say, I'm just . . . all these things that I have thought over because I've tried to introspect and I just follow one path and I say, "No, it couldn't be that." But like I say, it may be something that I don't want to face. It may be some psychological upheaval or doubt or question or something that I don't want to face. Like you said, sometimes others can help us see these things. We just get to the point where we could solve them, but the barrier which was built up against something that we don't ever want to face again won't permit us to solve the situation. It may . . . it may hark back to years past. And I was just hoping that through this, through your line of questioning, you might get a lead on something that I am trying to suppress.

Responses:

- sc1. These inner feelings that you have had about yourself, and that have existed over a period of time . . . you feel those are the contributing things that you have to overcome. Is that it?
- sc2. Well, if you remember a while back in one of your statements you

stated that maybe something outside of this particular thing . . . something in your earlier life that seems to be bothering . . . bothering you at this time.

- sc3. In other words, in trying to solve this problem for yourself, you seem to get in a rut and run up against a blank wall in trying to figure it out for yourself.
- sc4. Quite probably your problem is tied up in this suppression.
- sc5. You have a feeling that I might be able to assist you in uncovering some of this feeling that you have had.
- sc6. Well, what do *you* think it could be?
- sc7. Then you would like to remove those barriers that you mentioned. Is that it?
- sc8. Yes? (*Weakly.*)
- sc9. You've been doing some good thinking on this yourself, and probably with some more study you'll be able to arrive at a good conclusion.

CLIENT. That's right. That's the reason I'm inclined to . . . I believe it must be something like that, more so than anything else, it's the emotional thing. That I'm aware of. And the idea of the dreams—For there my conscious mind is not to the fore and the subconscious has more power and it keeps coming to the front—and possible, I mean, that's the reason that I think that possibly it might be something that I'm trying to hide from myself, I mean, that I don't want to face. What it is or what connection it has or when it happened or if it happened or what, I don't know.

Responses:

- sc1. I see. (*pause*) In other words your disturbances in your dreams all bring about this nervous tension inside yourself, and it is very difficult to release it properly. Is that it?
- sc2. Yes, it's very true. Many times we don't want to face the facts . . . and we go round . . . as is known . . . we go round the bush trying to find an outlet. But, sooner or later we will have to face the facts, and the sooner that we get down to the whole thing, perhaps we may be able to give you some insight, and also this friend of yours that you speak of.
- sc3. There seems to be some material deep down in . . . that comes up in your dreams, something that probably you are suppressing in your waking moments. Is that it?
- sc4. You seem to be really mixed up.
- sc5. You just don't know what might be behind all of this.

- sc6. You feel there is something there, yet you can't quite figure it out.
- sc7. Those dreams that you mentioned are really worrying you— isn't it?
- sc8. You feel, then, that this has been more or less an undesirable happening that you've had and you don't want to face it quite squarely.
- sc9. You're much further along in your thinking than many people. Many people won't face their problems as well as you've already done so.

CLIENT. Are we getting anywhere on that? I mean we plunged right into the thing. Do you want to know . . . I mean possibly you are learning about me, but I've just plunged right into it. You've got no tie-up with my background or anything. Possibly that comes out in time, but anyway I want to . . . am I getting you anywhere? This is all . . . I've gone over all this. I'm not getting anywhere. I'm just telling you things that I've gone over, see.

Responses:

- sc1. Any progress of talking about it is worth while for the individual. Once we establish something within you to better understand yourself. I'm sure you will fully realize it as well as I will. It's just for me to reflect what you're feeling. Now tell me more about these inner frustrations.
- sc2. If you would . . . uh—like to make an appointment for next Thursday afternoon at three o'clock, I'd like to see you in my office if it is possible.
- sc3. Uh—you're rather disappointed, then, in the results you expected to accomplish. You had really hoped that I would give you an answer, then, on the basis of these facts.
- sc4. If we can keep exploring for some time longer, perhaps we can—uh—uncover some new material.
- sc5. You feel that you just keep repeating this over and over to me and wondering what I think about it.
- sc6. Perhaps just verbalizing your feeling will help you see the problem more clearly and present a solution.
- sc7. Would you like to—uh—tell me something more about those dreams at this time?
- sc8. Yes—but you are interested, very interested in solving your own problem as well as you can.
- sc9. It's altogether possible, though, that in going over these things that each time you are understanding some things about yourself that you haven't understood before.

CLIENT. That's right. So that . . . That's right, what you said there . . . I'm asking you what the best way, whether you want to plunge right into it or whether you want to learn about me or my interest or my . . . well, whatever it is, so that you can help me. Miss _____ asked me . . . she kind of . . . when I was talking with her about this, she said one of the main things is that you're going to have to be cooperative. Well, I've tried to allay any doubts on her part. This is one thing that I've based everything on. Because, as I told her, I wanted to go to a psychiatrist or psychologist, but the fees are prohibitive. It was just beyond me. I want to continue my education and that's what I want. So right now I couldn't do that. But this thing is a block, it seems to me, to my education. So I told her that I would be willing to do all I could, and she said, "Well, with that attitude you ought to see a counselor."

Responses:

- sc1. Actually you mean to say that education, now, after we have gotten through this far in the counseling session you can see that it holds for you what you need. Is that what you mean?
- sc2. Well, as I've made an appointment for you with the psychiatrist, and after he has had an opportunity of talking with you and giving these tests, then we may be able to derive a better solution, if it is possible. We'll do all we possibly can, and if it is satisfactory with you, I'll put you down here for three o'clock next Thursday afternoon.
- sc3. It means a lot to you, then, accomplishing what you want to in this field.
- sc4. You can feel free to talk about anything, because we'll certainly keep it confidential and if you'd like to plunge right into this thing which is bothering you most, we can do some good probably.
- sc5. So you came because she suggested that you come . . . or did you come—(no wait a minute, I'm too far off base). Suppose we consider this thing just a little bit further. I believe you had something in mind in coming over here. You didn't come just because she sent you over. Won't you go just a bit further with this, and let's see just what we might be able to work out?
- sc6. You're really willing to try to solve the problem.
- sc7. The question is not whether you are getting me anywhere, but whether or not you are getting yourself somewhere. Now, if you would like to discuss this at further length I'd be glad to discuss it with you. Sometimes we don't know when we are progressing until we sit down and think things over. Then you feel that now you are

here seeing the counselor you would like to work something out, so you would know which step to take next. Is that it?

- sc8. You feel, then, that by getting more education, that this will largely overcome your problems.
- sc9. It certainly seems to me that in time you're going to get the solution of your problem. As I said before not many people will allow themselves to think that they do have problems.

EXAMPLE OF PERFORMANCE AND SCORING OF A NONDIRECTIVE PERFORMANCE TEST

COMPARATIVE interviews upon a performance test of student counselors were rated independently by professional psychologists. Each interview, complete within itself, contains the client's statements and the student counselor's statements. That is, the same client made the same statements to each student counselor in turn. See Appendix B for cross comparison of the responses of nine different student counselors to each of the successive statements made by the same "client." Comparative performance is indicated.

1. The interview which was rated highest (SC5):

CLIENT. I would like to be a psychologist. I understand you're a psychologist. I'm working in the daytime, I'm going to school at night. And right now it looks like that's what my whole life is wrapped up in education. That is the sole thing I want. And I want to find out whether I'm capable, whether I'm mentally capable. I seem to be, but I'm not sure. I have a doubt, and Mrs. —, she was my general science, general psychology, yes, my general psychology, instructor, she gave us some type of IQ. I don't know exactly what type it was, but it did prove very favorable. But she said it was just an experiment, nothing for me to take definitely. Well, I would like to have some kind of definite understanding there of my own capabilities. I did very well in high school and I seem to be doing very well in college, but with all the money I'm going to have to spend and the hopes I'm putting in my future through education, I would like to know just where I stand.

SC. You really want to find out what you—uh—what might be best for you. Is that right?

CLIENT. The kind or type of work I want to go into is child guidance work. From what I understand, all this is just through what I have asked questions about and so forth. I have no facts or anything. But I talked to Father _____ about it and around the Center and the YMCA, etc. There are not very many Negroes in that specific type of work—child guidance—and I do have that handicap to come over, to get into the field to break down the barriers. I don't mind all of that if I know that I can get somewhere. I don't want—I want a livelihood but I'm not in it to have a big house and a lot of money and everything. I want to go into it because I think I like it.

SC. You feel that you would like child psychology?

CLIENT. That's right. I want to be confident in myself and it doesn't matter what I have to face.

SC. You're not so sure.

CLIENT. That's right. Because I mean it's not . . . it's a sacrifice now.

I work in a factory, and, well, with working and studying until two o'clock in the morning, and my physical condition is good, but I am extremely nervous and it does slightly affect my heart and I did have ulcers when I was eighteen and . . . I never had operation, but I had a milk diet and all that which never really cured them up. All of that, my mental situation and my physical situation, my nerves and everything, I want to know that I can do it on my own regardless of the obstacles if I have confidence in myself. So I'd like to have some type of IQ, if you can give it. I don't have any money, but whatever the charge is, I'll try to meet it. Miss _____ at the time told me they were very expensive, but she gave me this one free. Took about five hours one evening, and she said it was very good. But that was all. She didn't tell me just where I was weak or what I was good in and so forth.

SC. I see-e-e. (*Deliberately.*)

CLIENT. M-hm. Well, I wonder if we had the results of the test that Miss _____ gave me, if we had them available during the counseling, could we derive benefit from them? Could we use them . . . could they possibly do a lot to clear up some doubts?

SC. You're just not for sure whether you are physically able to move into this type of a training program.

CLIENT. Yes, helpful to me, but along with what you are doing. Because I don't know, I mean maybe, maybe, but to me the whole thing is tied up together, all this, the emotional and the intellectual may be

that together, maybe not, I don't know. But you can appreciate my position when you realize that on my education, I feel like that's my whole future. I don't definitely, I definitely do not want to work in a factory all my life. I mean, I believe I could go insane doing that. And there seems to be nothing else in life for me but that. I mean I do want a home and a family and so forth, but to have a home and family I've got to have something else too. And it seems, everything now seems to be education, but there seems to be some emotional or psychological obstacle or barrier or frustration. And I was wondering if somewhere that is tied in also. That's not the main thing, that's just something I was wondering, you see, and, and if possible could we use the IQ test to possible open up a road to what really *is* bothering me emotionally.

SC. Do you feel that the test you have had might be of some value to you in this situation?

CLIENT. That's right. That's what I told Miss _____ when I talked to her.

SC. You feel like the results of the intelligence tests you have had might be of some help to you in reaching a definite conclusion.

CLIENT. That's it. The whole thing, I am rather convinced it's an emotional setup, because I'm nervous, and I dream at night, I'm exhausted and I really sleep; I mean, I'm very sleepy all the time though I try to get eight to nine hours' sleep at night. Some nights I can't because of school, but other nights then I'll get maybe ten or eleven hours, something like that. But somehow I try to get enough sleep. But I dream constantly. Crazy dreams that I know must have as a . . . as I said, I'm going to major in psychology some day, so I'm . . . I do a little bit of reading in psychology. Possibly it is wrong, possibly because of my lack of knowledge I'm drawing the wrong conclusions, but I believe these dreams are the results of some emotional stress.

SC. I see.

CLIENT. As we talk, I'll be able to bring them out more clearly, but usually when I'm in an emotional upheaval, I have more of these dreams. Just a period about, oh, about three weeks ago, up to that period for about six weeks, well, I had final exams at school and everything was going along placidly, though it was final exams at school and everything and I made B's and A's in my grades and so forth. But then all of a sudden a disturbance came in my life and it was presaged by a period of dreaming, even before I knew what was going to happen. That made me sensitive to some type of trouble. And the greatest thing that happened in my life happened to me, detrimentally,

and, well, all that, it just doesn't tie up or it does tie up, I don't know what.

SC. You feel that the dreams you are having might have some connection with the emotional upset that you are worrying about. Is that the way you are feeling about it?

CLIENT. Well, now, I mean, I'm about talked out. (*Laughs.*) I mean I could talk more but that was by way of introduction, so that you could find out just what I want.

SC. You're not so sure whether there is any connection between this disturbance and your dreams.

CLIENT. That's right.

SC. Well, perhaps if we can go into this just a little bit further, perhaps we can reach some definite conclusion.

CLIENT. That's right. That's right.

SC. Well, tell me, tell me just a little bit more about this situation.

CLIENT. I don't know just what causes it. I mean, it may be something in my personal life, or it may be something that happened to me that I have tried to talk in my mind and bury away, and subconsciously I keep holding it down, but it wants to come to the front. And I don't know what it is, and I wish I could find what it is and solve it once and for all, and everything would dissolve. I don't know whether it's financial or, I mean, I don't know what it is. But it's something. I mean it's *got* to be something.

SC. Even though you try not to think about it, it still develops sooner or later.

CLIENT. That's right. I have a friend (*pause*), well, anyway, he and I do a lot of talking. And there's something that we have a mutual agreement on. I just don't know what it is. But I just met him about three years ago. And our friendship has increased. We enjoy sitting down and talking to each other. He has the same type of problem. We just don't know; we tried to sit down and just figure out what it was. He wants an education and he's not getting anywhere. I want an education. It looks like I'm getting the education, but I don't seem to be getting anywhere, even though I'm getting an education. Maybe that's just a belief of my own. Possible. But I'm twenty-seven and all I have is a car and working for an education.

SC. You seem to be headed in some special direction but you just don't know where you're going.

CLIENT. That's right. And I believe that maybe it's because I doubt whether this is the path . . . I believe it is . . . doubt whether this

is the path or whether I'm capable of doing what I want to do when I am going through this way or what, I don't know. That may be the doubt or as I say, it may be something else in my life. Really, I don't know what it is. And if maybe just a big (*pause*) but I don't think so, because it affects me too much. I make friends and I'm always at odds when I don't see things their way, so consequently, I mean my friendships are temporary. And everybody tells me I'm radical, I'm, I'm idealistic all that. And I don't know whether this one fellow and I remained friends because we have a common problem or a common understanding.

sc. You have a feeling that your friend understands the way you feel about these things.

CLIENT. Well, I believe it *is* education. I mean, I don't see right now anything else I want. What type of education . . . Am I in the right education? I know it's not science, because I'm no good at math, or anything like that. In philosophy in school I excel. In psychology and sociology too. So it must be something along the human interest field.

sc. You have a feeling that your education is absolutely necessary to give you the desires that you have in life.

CLIENT. That's right. Like I say, I'm just . . . all these things that I have thought over because I've tried to introspect and I just follow one path and I say, "No, it couldn't be that." But like I say, it may be something that I don't want to face. It may be some psychological upheaval or doubt or question or something that I don't want to face. Like you said, sometimes others can help us see these things. We just get to the point where we could solve them, but the barrier which was built up against something that we don't ever want to face again won't permit us to solve the situation. It may . . . it may hark back to years past. And I was just hoping that through this, through your line of questioning, you might get a lead on something that I am trying to suppress.

sc. You have a feeling that I might be able to assist you in uncovering some of this feeling that you have had.

CLIENT. That's right. That's the reason I'm inclined to . . . I believe it must be something like that, more so than anything else, it's the emotional thing. That I'm aware of. And the idea of the dreams— For there my conscious mind is not to the fore and the subconscious has more power and it keeps coming to the front—and possible, I mean, that's the reason that I think that possibly it might be something that I'm trying to hide from myself, I mean, that I don't want to face.

What it is or what connection it has or when it happened or if it happened or what, I don't know.

sc. You just don't know what might be behind all of this.

CLIENT. Are we getting anywhere on that? I mean we plunged right into the thing. Do you want to know . . . I mean possibly you are learning about me, but I've just plunged right into it. You've got no tie-up with my background or anything. Possibly that comes out in time, but anyway I want to . . . am I getting you anywhere? This is all . . . I've gone over all this. I'm not getting anywhere. I'm just telling you things that I've gone over, see.

sc. You feel that you just keep repeating this over and over to me and wondering what I think about it.

CLIENT. That's right. So that . . . That's right, what you said there . . . I'm asking you what the best way, whether you want to plunge right into it or whether you want to learn about me or my interest or my . . . well, whatever it is, so that you can help me. Miss _____ asked me . . . she kind of . . . when I was talking with her about this, she said one of the main things is that you're going to have to be cooperative. Well, I've tried to allay any doubts on her part. This is one thing that I've based everything on. Because, as I told her, I wanted to go to a psychiatrist or psychologist, but the fees are prohibitive. It was just beyond me. I wanted to continue my education and that's what I want. So right now I couldn't do that. But this thing is a block, it seems to me, to my education. So I told her that I would be willing to do all I could, and she said, "Well, with that attitude you ought to see a counselor."

sc. So you came because she suggested that you come . . . or did you come—(no wait a minute, I'm too far off base). Suppose we consider this thing just a little bit further. I believe you had something in mind in coming over here. You didn't just come because she sent you over. Won't you go just a bit further with this, and let's see just what we might be able to work out?

2. The interview which was rated lowest (SC2):

CLIENT. I would like to be a psychologist. I understand you're a psychologist. I'm working in the daytime, I'm going to school at night. And right now it looks like that's what my whole life is wrapped up in education. That is the sole thing I want. And I want to find out whether I'm capable, whether I'm mentally capable. I seem to be, but I'm not sure. I have a doubt, and Mrs. _____, she was my general

science, general psychology, yes, my general psychology, instructor, she gave us some type of IQ. I don't know exactly what type it was, but it did prove very favorable. But she said it was just an experiment, nothing for me to take definitely. Well, I would like to have some kind of definite understanding there of my own capabilities. I did very well in high school and I seem to be doing very well in college, but with all the money I'm going to have to spend and the hopes I'm putting in my future through education, I would like to know just where I stand.

sc. No, I may inform you that I am not a psychologist. As you have stated that you would like some sort of testing to see just where you stand. There are many different types of tests that perhaps we may give you—before we go much further in this interview.

CLIENT. The kind or type of work I want to go into is child guidance work. From what I understand, all this is just through what I have asked questions about and so forth. I have no facts or anything. But I talked to Father _____ about it and around the Center and the YMCA, etc. There are not very many Negroes in that specific type of work—child guidance—and I do have that handicap to come over, to get into the field to break down the barriers. I don't mind all of that if I know that I can get somewhere. I don't want—I want a livelihood but I'm not in it to have a big house and a lot of money and everything. I want to go into it because I think I like it.

sc. You lack a particular background for this type of work, outside of the fact that you feel you would like to go into it for a livelihood.

CLIENT. That's right. I want to be confident in myself and it doesn't matter what I have to face.

sc. (*Pause—long pause.*)

CLIENT. That's right. Because I mean it's not . . . it's a sacrifice now. I work in a factory, and, well, with working and studying until two o'clock in the morning, and my physical condition is good, but I am extremely nervous and it does slightly affect my heart and I did have ulcers when I was eighteen and . . . I never had operation, but I had a milk diet and all that which never really cured them up. All of that, my mental situation and my physical situation, my nerves and everything, I want to know that I can do it on my own regardless of the obstacles if I have confidence in myself. So I'd like to have some type of IQ, if you can give it. I don't have any money, but whatever the charge is, I'll try to meet it. Miss _____ at the time told me they were very expensive, but she gave me this one free. Took about five hours one evening, and she said it was very good. But that

was all. She didn't tell me just where I was weak or what I was good in and so forth.

SC. First, I think that we should give you some more tests, to see what particular thing that you are best fitted for. You may not even—you may be fitted for something else outside the field of child guidance that you seem to be so interested in.

CLIENT. M-hm. Well, I wonder if we had the results of the test that Miss _____ gave me, if we had them available during the counseling, could we derive benefit from them? Could we use them . . . could they possibly do a lot to clear up some doubts?

SC. What makes you particularly think that you would be good in this type of work even though—that there are only a few colored people that follow the child guidance program?

CLIENT. Yes, helpful to me, but along with what you are doing. Because I don't know, I mean maybe, maybe, but to me the whole thing is tied up together, all this, the emotional and the intellectual may be that together, maybe not, I don't know. But you can appreciate my position when you realize that on my education, I feel like that's my whole future. I don't definitely, I definitely do not want to work in a factory all my life. I mean, I believe I could go insane doing that. And there seems to be nothing else in life for me but that. I mean I do want a home and a family and so forth, but to have a home and family I've got to have something else too. And it seems, everything now seems to be education, but there seems to be some emotional or psychological obstacle or barrier or frustration. And I was wondering if somewhere that is tied in also. That's not the main thing, that's just something I was wondering, you see, and, and if possible could we use the IQ test to possible open up a road to what really *is* bothering me emotionally.

SC. That depends upon the type of test that she gave you and I would not be stating that they would absolutely give us the correct answer.

CLIENT. That's right. That's what I told Miss _____ when I talked to her.

SC. It may be of some help to us in this particular solution. I wouldn't say altogether.

CLIENT. That's it. The whole thing, I am rather convinced it's an emotional setup, because I'm nervous, and I dream at night, I'm exhausted and I really sleep; I mean, I'm very sleepy all the time though I try to get eight to nine hours' sleep at night. Some nights I can't because of school, but other nights then I'll get maybe ten or eleven hours, something like that. But somehow I try to get enough sleep. But I dream constantly. Crazy dreams that I know must have as a . . . as I said,

I'm going to major in psychology some day, so I'm . . . I do a little bit of reading in psychology. Possibly it is wrong, possibly because of my lack of knowledge I'm drawing the wrong conclusions, but I believe these dreams are the results of some emotional stress.

SC. Well, now just tell me, on this test that Miss _____ gave to you—what did she tell you about this test? Give me a little more of it.

CLIENT. As we talk, I'll be able to bring them out more clearly, but usually when I'm in an emotional upheaval, I have more of these dreams. Just a period about, oh, about three weeks ago, up to that period for about six weeks, well, I had final exams at school and everything was going along placidly, though it was final exams at school and everything and I made B's and A's in my grades and so forth. But then all of a sudden a disturbance came in my life and it was presaged by a period of dreaming, even before I knew what was going to happen. That made me sensitive to some type of trouble. And the greatest thing that happened in my life happened to me, detrimentally, and, well, all that, it just doesn't tie up or it does tie up, I don't know what.

SC. Perhaps we should refer you to one of the psychiatrists here in the building. If you have a few minutes, I'll try to make an appointment for you.

CLIENT. Well, now, I mean, I'm about talked out. (*Laughs.*) I mean I could talk more but that was by way of introduction, so that you could find out just what I want.

SC. From what you tell me it seems as that there is something else here that seems to be bothering you outside of this particular degree or the work that you want to get in child guidance. Now we may be able to get down here to a final answer if you just come out and put everything on the table.

CLIENT. That's right.

SC. Just a minute, I'll call Dr. X and make an appointment for you and you can go on down and see what he has to talk to you about.

CLIENT. That's right. That's right.

SC. He—uh—he's out at present, right at this particular moment and the girl said it would be all right for you to come right on down there.

CLIENT. I don't know just what causes it. I mean, it may be something in my personal life, or it may be something that happened to me that I have tried to talk in my mind and bury away, and subconsciously I keep holding it down, but it wants to come to the front. And I don't know what it is, and I wish I could find what it is and solve it once and for all, and everything would dissolve. I don't know whether

it's financial or, I mean, I don't know what it is. But it's something. I mean it's *got* to be something.

SC. Yes, there must be an answer to all questions. And it is now about time that the doctor should be back in his office, so we'll just go right on down.

CLIENT. That's right. I have a friend (*pause*), well, anyway, he and I do a lot of talking. And there's something that we have a mutual agreement on. I just don't know what it is. But I just met him about three years ago. And our friendship has increased. We enjoy sitting down and talking to each other. He has the same type of problem. We just don't know; we tried to sit down and just figure out what it was. He wants an education and he's not getting anywhere. I want an education. It looks like I'm getting the education, but I don't seem to be getting anywhere, even though I'm getting an education. Maybe that's just a belief of my own. Possible. But I'm twenty-seven and all I have is a car and working for an education.

SC. Well, after you have had a chance to talk to Dr. X, perhaps he'll be able to enlighten you on some of these personal problems that we haven't gone into, that seem to be bothering you.

CLIENT. That's right. And I believe that maybe it's because I doubt whether this is the path . . . I believe it is . . . doubt whether this is the path or whether I'm capable of doing what I want to do when I am going through this way or what, I don't know. That may be the doubt or as I say, it may be something else in my life. Really, I don't know what it is. And if maybe just a big (*pause*) but I don't think so, because it affects me too much. I make friends and I'm always at odds when I don't see things their way, so consequently, I mean my friendships are temporary. And everybody tells me I'm radical, I'm, I'm idealistic all that. And I don't know whether this one fellow and I remain friends because we have a common problem or a common understanding.

SC. There seems to be some doubt in your mind, about this particular work that you spoke about . . . this child guidance . . . and whether you would be able to go on with the work if you did have the proper qualifications.

CLIENT. Well, I believe it *is* education. I mean, I don't see right now anything else I want. What type of education . . . am I in the right education? I know it's not science, because I'm no good at math, or anything like that. In philosophy in school I excel. In psychology and sociology too. So it must be something along the human interest field.

SC. After you have taken these tests that we are going to give you, the

final results will give us something to work on as to whether you are qualified to go into this type of work that you seem to be so interested in.

CLIENT. That's right. Like I say, I'm just . . . all these things that I have thought over because I've tried to introspect and I just follow one path and I say, "No, it couldn't be that." But like I say, it may be something that I don't want to face. It may be some psychological upheaval or doubt or question or something that I don't want to face. Like you said, sometimes others can help us see these things. We just get to the point where we could solve them, but the barrier which was built up against something that we don't ever want to face again won't permit us to solve the situation. It may . . . it may hark back to years past. And I was just hoping that through this, through your line of questioning, you might get a lead on something that I am trying to suppress.

sc. Well, if you remember a while back in one of your statements you stated that maybe something outside of this particular thing . . . something in your earlier life that seems to be bothering . . . bothering you at this time.

CLIENT. That's right. That's the reason I'm inclined to . . . I believe it must be something like that, more so than anything else, it's the emotional thing. That I'm aware of. And the idea of the dreams—For there my conscious mind is not to the fore and the subconscious has more power and it keeps coming to the front—and possible, I mean, that's the reason that I think that possibly it might be something that I'm trying to hide from myself, I mean, that I don't want to face. What it is or what connection it has or when it happened or if it happened or what, I don't know.

sc. Yes, it's very true. Many times we don't want to face the facts . . . and we go round . . . as is known . . . we go round the bush trying to find an outlet. But, sooner or later we will have to face the facts, and the sooner that we get down to the whole thing, perhaps we may be able to give you some insight, and also this friend of yours that you speak of.

CLIENT. Are we getting anywhere on that? I mean we plunged right into the thing. Do you want to know . . . I mean possibly you are learning about me, but I've just plunged right into it. You've got no tie-up with my background or anything. Possibly that comes out in time, but anyway I want to . . . am I getting you anywhere? This is all . . . I've gone over all this. I'm not getting anywhere. I'm just telling you things that I've gone over, see.

sc. If you would . . . uh—like to make an appointment for next Thurs-

day afternoon at three o'clock, I'd like to see you in my office if it is possible.

CLIENT. That's right. So that . . . that's right, what you said there . . . I'm asking you what the best way, whether you want to plunge right into it or whether you want to learn about me or my interest or my . . . well, whatever it is, so that you can help me. Miss _____ asked me . . . she kind of . . . when I was talking with her about this, she said one of the main things is that you're going to have to be co-operative. Well, I've tried to allay any doubts on her part. This is one thing that I've based everything on. Because, as I told her, I wanted to go to a psychiatrist or psychologist, but the fees are prohibitive. It was just beyond me. I want to continue my education and that's what I want. So right now I couldn't do that. But this thing is a block, it seems to me, to my education. So I told her that I would be willing to do all I could, and she said, "Well, with that attitude you ought to see a counselor."

SC. Well, as I've made an appointment for you with the psychiatrist, and after he has had an opportunity of talking with you and giving these tests, then we may be able to derive a better solution, if it is possible. We'll do all we possibly can, and if it is satisfactory with you, I'll put you down here for three o'clock next Thursday afternoon.

3. The typical or median-quality interview (SC8):

CLIENT. I would like to be a psychologist. I understand you're a psychologist. I'm working in the daytime, I'm going to school at night. And right now it looks like that's what my whole life is wrapped up in education. That is the sole thing I want. And I want to find out whether I'm capable, whether I'm mentally capable. I seem to be, but I'm not sure. I have a doubt, and Mrs. _____, she was my general science, general psychology, yes, my general psychology, instructor, she gave us some type of IQ. I don't know exactly what type it was, but it did prove very favorable. But she said it was just an experiment, nothing for me to take definitely. Well, I would like to have some kind of definite understanding there of my own capabilities. I did very well in high school and I seem to be doing very well in college, but with all the money I'm going to have to spend and the hopes I'm putting in my future through education, I would like to know just where I stand.

SC. You're a little unsure as to exactly just what you want to do. Is that right?

CLIENT. The kind or type of work I want to go into is child guidance work.

From what I understand, all this is just through what I have asked questions about and so forth. I have no facts or anything. But I talked to Father _____ about it and around the Center and the YMCA, etc. There are not very many Negroes in that specific type of work—child guidance—and I do have that handicap to come over, to get into the field to break down the barriers. I don't mind all of that if I know that I can get somewhere. I don't want—I want a livelihood but I'm not in it to have a big house and a lot of money and everything. I want to go into it because I think I like it.

SC. You're really interested in that kind of work.

CLIENT. That's right. I want to be confident in myself and it doesn't matter what I have to face.

SC. Yes? (*Weakly.*)

CLIENT. That's right. Because I mean it's not . . . it's a sacrifice now.

I work in a factory, and, well, with working and studying until two o'clock in the morning, and my physical condition is good, but I am extremely nervous and it does slightly affect my heart and I did have ulcers when I was eighteen and . . . I never had operation, but I had a milk diet and all that which never really cured them up. All of that, my mental situation and my physical situation, my nerves and everything, I want to know that I can do it on my own regardless of the obstacles if I have confidence in myself. So I'd like to have some type of IQ, if you can give it. I don't have any money, but whatever the charge is, I'll try to meet it. Miss _____ at the time told me they were very expensive, but she gave me this one free. Took about five hours one evening, and she said it was very good. But that was all. She didn't tell me just where I was weak or what I was good in and so forth.

SC. You're—you like to know just where you stand.

CLIENT. M-hm. Well, I wonder if we had the results of the test that Miss _____ gave me, if we had them available during the counseling, could we derive benefit from them? Could we use them . . . could they possibly do a lot to clear up some doubts?

SC. They may. We'll look that up later.

CLIENT. Yes, helpful to me, but along with what you are doing. Because I don't know, I mean maybe, maybe, but to me the whole thing is tied up together, all this, the emotional and the intellectual may be that together, maybe not, I don't know. But you can appreciate my position when you realize that on my education, I feel like that's my whole future. I don't definitely, I definitely do not want to work in a factory all my life. I mean, I believe I could go insane doing that.

And there seems to be nothing else in life for me but that. I mean I do want a home and a family and so forth, but to have a home and family I've got to have something else too. And it seems, everything now seems to be education, but there seems to be some emotional or psychological obstacle or barrier or frustration. And I was wondering if somewhere that is tied in also. That's not the main thing, that's just something I was wondering, you see, and, and if possible could we use the IQ test to possible open up a road to what really is bothering me emotionally.

SC. You're wondering if the IQ tests will give an indication of your emotional problems.

CLIENT. That's right. That's what I told Miss _____ when I talked to her.

SC. Yes. (*Weakly.*)

CLIENT. That's it. The whole thing, I am rather convinced it's an emotional setup, because I'm nervous, and I dream at night, I'm exhausted and I really sleep; I mean, I'm very sleepy all the time though I try to get eight to nine hours' sleep at night. Some nights I can't because of school, but other nights then I'll get maybe ten or eleven hours, something like that. But somehow I try to get enough sleep. But I dream constantly. Crazy dreams that I know must have as a . . . as I said, I'm going to major in psychology some day, so I'm . . . I do a little bit of reading in psychology. Possibly it is wrong, possibly because of my lack of knowledge I'm drawing the wrong conclusions, but I believe these dreams are the results of some emotional stress.

SC. You're interested then in the connection between your dreams and psychology as you like to study it?

CLIENT. As we talk, I'll be able to bring them out more clearly, but usually when I'm in an emotional upheaval, I have more of these dreams. Just a period about, oh, about three weeks ago, up to that period for about six weeks, well, I had final exams at school and everything was going along placidly, though it was final exams at school and everything and I made B's and A's in my grades and so forth. But then all of a sudden a disturbance came in my life and it was presaged by a period of dreaming, even before I knew what was going to happen. That made me sensitive to some type of trouble. And the greatest thing that happened in my life happened to me, detrimentally, and, well, all that, it just doesn't tie up or it does tie up, I don't know what.

SC. You believe then there is some connection between your dreams and your problems as they occur?

CLIENT. Well, now, I mean, I'm about talked out. (*Laughs.*) I mean I

could talk more but that was by way of introduction, so that you could find out just what I want.

SC. H-um, I see.

CLIENT. That's right.

SC. (*Paused and blocked.*)

CLIENT. That's right. That's right.

SC. (*Paused and blocked.*)

CLIENT. I don't know just what causes it. I mean, it may be something in my personal life, or it may be something that happened to me that I have tried to talk in my mind and bury away, and subconsciously I keep holding it down, but it wants to come to the front. And I don't know what it is, and I wish I could find what it is and solve it once and for all, and everything would dissolve. I don't know whether it's financial or, I mean, I don't know what it is. But it's something. I mean it's *got* to be something.

SC. You feel that your problem is almost an obvious one.

CLIENT. That's right. I have a friend (*pause*), well, anyway, he and I do a lot of talking. And there's something that we have a mutual agreement on. I just don't know what it is. But I just met him about three years ago. And our friendship has increased. We enjoy sitting down and talking to each other. He has the same type of problem. We just don't know; we tried to sit down and just figure out what it was. He wants an education and he's not getting anywhere. I want an education. It looks like I'm getting the education, but I don't seem to be getting anywhere, even though I'm getting an education. Maybe that's just a belief of my own. Possible. But I'm twenty-seven and all I have is a car and working for an education.

SC. You're interested then, in talking with your friend in trying to solve your problems.

CLIENT. That's right. And I believe that maybe it's because I doubt whether this is the path . . . I believe it is . . . doubt whether this is the path or whether I'm capable of doing what I want to do when I am going through this way or what, I don't know. That may be the doubt or as I say, it may be something else in my life. Really, I don't know what it is. And if maybe just a big (*pause*) but I don't think so, because it affects me too much. I make friends and I'm always at odds when I don't see things their way, so consequently, I mean my friendships are temporary. And everybody tells me I'm radical, I'm, I'm idealistic all that. And I don't know whether this one fellow and I remained friends because we have a common problem or a common understanding.

SC. You're interested in keeping up your friends that you meet and you do make friends easily but this friend of yours still seems to be the one you like to talk to and understand the most.

CLIENT. Well, I believe it *is* education. I mean, I don't see right now anything else I want. What type of education . . . Am I in the right education? I know it's not science, because I'm no good at math, or anything like that. In philosophy in school I excel. In psychology and sociology too. So it must be something along the human interest field.

SC. You are, then, mostly interested in people as you meet them.

CLIENT. That's right. Like I say, I'm just . . . all these things that I have thought over because I've tried to introspect and I just follow one path and I say, "No, it couldn't be that." But like I say, it may be something that I don't want to face. It may be some psychological upheaval or doubt or question or something that I don't want to face. Like you said, sometimes others can help us see these things. We just get to the point where we could solve them, but the barrier which was built up against something that we don't ever want to face again won't permit us to solve the situation. It may . . . it may hark back to years past. And I was just hoping that through this, through your line of questioning, you might get a lead on something that I am trying to suppress.

SC. Yes? (*Weakly.*)

CLIENT. That's right. That's the reason I'm inclined to . . . I believe it must be something like that, more so than anything else, it's the emotional thing. That I'm aware of. And the idea of the dreams—For there my conscious mind is not to the fore and the subconscious has more power and it keeps coming to the front—and possible, I mean, that's the reason that I think that possibly it might be something that I'm trying to hide from myself, I mean, that I don't want to face. What it is or what connection it has or when it happened or if it happened or what, I don't know.

SC. You feel, then, that this has been more or less an undesirable happening that you've had and you don't want to face it quite squarely.

CLIENT. Are we getting anywhere on that? I mean we plunged right into the thing. Do you want to know . . . I mean possibly you are learning about me, but I've just plunged right into it. You've got no tie-up with my background or anything. Possibly that comes out in time, but anyway I want to . . . am I getting you anywhere? This is all . . . I've gone over all this. I'm not getting anywhere. I'm just telling you things that I've gone over, see.

sc. Yes—but you are interested, very interested in solving your own problem as well as you can.

CLIENT. That's right. So that . . . That's right, what you said there . . . I'm asking you what the best way, whether you want to plunge right into it or whether you want to learn about me or my interest or my . . . well, whatever it is, so that you can help me. Miss _____ asked me . . . she kind of . . . when I was talking with her about this, she said one of the main things is that you're going to have to be co-operative. Well, I've tried to allay any doubts on her part. This is one thing that I've based everything on. Because, as I told her, I wanted to go to a psychiatrist or psychologist, but the fees are prohibitive. It was just beyond me. I want to continue my education and that's what I want. So right now I couldn't do that. But this thing is a block, it seems to me, to my education. So I told her that I would be willing to do all I could, and she said, "Well, with that attitude you ought to see a counselor."

sc. You feel, then, that by getting more education, that this will largely overcome your problems.

The following scale is quoted from Stone.¹

Techniques in the Non-directive, Directive Counseling Continuum (From Non-directive to Directive).

1. Non-directive Listening
 - a. Head nodding and shaking (Non-directive)
 - b. Gesture and facial expression (Non-directive)
2. Um-hm (Non-directive)
3. "Reflecting"
 - a. Repeating the last few words, or using synonyms to keep current thought in motion
 - b. Reflection of basic feeling
4. Clarification (Rearrangement, Restatement in Different Words)
5. Selection of Part
6. Fact Giving (Simple)
 - a. For simple direction (including structuring a relationship)
 - b. For reassurance
 - c. Directive head shaking, nodding, and "Um-hm," and
 - d. Directive gestures and facial expression
7. Fact Getting
 - a. From records, and

¹ Stone, D. R., "Logical Analysis of the Directive, Non-directive Counseling Continuum," *Voc. Guidance J.*, February, 1950.

- b. From the individual by interview question, and
 - c. From testing
- 8. Fact Giving (Complex)
 - a. Diagnosing, analyzing, interpreting, synthesizing
- 9. Prognosis
- 10. Directed Problem Solving
- 11. Tempered Command (You Should)
- 12. Command (You Must)
- 13. Changing the Environment
- 14. Force
 - a. Threat
 - b. Physical

COLLATERAL READINGS

- Adkins, Dorothy C. "Test Construction in Public Personnel Administration." *Educ. Psychol. Measmt.*, 1944, 4:141-160.
- . "Construction and Analysis of Written Tests for Predicting Job Performance." *Educ. Psychol. Measmt.*, 1946, 6:195-212.
- Anderson, J. E. "The Effect of Item Analysis upon the Discriminating Power of an Examination." *J. Appl. Psychol.*, 1935, 19:237-244.
- Anderson, R. G. "Test Scores and Efficiency Ratings of Machinists." *J. Appl. Psychol.*, 1947, 31:377-388.
- Barry, R. F. "An Analysis of Some New Statistical Methods of Selecting Test Items." *J. Exp. Educ.*, 1939, 7:221-228.
- Bransford, T. L., et al. "A Study of the Validity of Written Tests for Administrative Personnel." *Amer. Psychol.*, 1946, 7:279 (abstract).
- Burt, Cyril. "Validating Tests for Personnel Selection." *Brit. J. Psychol.*, 1943, 34:1-19.
- Carroll, J. B. "The Effect of Difficulty and Chance Success on Correlations between Items or between Tests." *Psychometrika*, 1945, 10:1-19.
- Cronbach, Lee J. "Response Sets and Test Validity." *Educ. Psychol. Measmt.*, 1946, 6:475-494.
- . "Further Evidence on Response Sets and Test Design." *Educ. Psychol. Measmt.*, 1950, 10:3-31.
- Crooks, W. R., and Ferguson, L. W. "Item Validities of Otis Self-administering Tests of Mental Ability for Colleges." *J. Exp. Educ.*, 1941, 9:229-232.
- Davidson, W. M., and Carroll, J. B. "Speed and Level Components in Time Limit Scores." *Educ. Psychol. Measmt.*, 1945, 5:411-428.
- Davis, F. B. *Item Analysis*. Cambridge, Mass.: Harvard University Press, 1947.
- Drake, R. M. "Factor Analysis of Music Tests." *Psychol. Bull.*, 1939, 26:608-609.

- Engelhart, M. D. "Unique Types of Achievement Test Exercises." *Psychometrika*, 1942, 7:103-115.
- . "Suggestions for Writing Achievement Exercises to Be Used in Tests Scored on the Electric Scoring Machine." *Educ. Psychol. Measmt.*, 1947, 7:351-374.
- Fiske, D. W. "Validation of Naval Aviation Cadet Selection Tests against Training Criteria." *J. Appl. Psychol.*, 1947, 31:601-614.
- Flanagan, J. C. "A Short Method of Selecting the Best Combination of Test Items for a Particular Purpose." *Psychol. Bull.*, 1936, 33:603-604.
- Forlano, G., and Pintner, R. "Selection of Upper and Lower Groups for Item Validation." *J. Educ. Psychol.*, 1941, 32:544-549.
- Garrett, H. E. *Statistics in Psychology and Education*, 3d ed. New York: Longmans, Green & Co., Inc., 1947.
- Gibbons, C. C. "The Predictive Value of the Most Valid Items of an Examination." *J. Educ. Psychol.*, 1940, 31:616-621.
- Green, H. A., and Jorgensen, A. N. *Measurement and Evaluation in the Secondary School*. New York: Longmans, Green & Co., Inc., 1943.
- Greene, Edward B. *Measurements of Human Behavior*. New York: The Odyssey Press, Inc., 1941.
- Guilford, J. P. "Factor Analysis in a Test Development Program." *Psychol. Rev.*, 1948, 35:79-94.
- . *Fundamental Statistics in Psychology and Education*, 2d ed. New York: McGraw-Hill Book Company, Inc., 1950.
- Gulliksen, Harold. *Theory of Mental Tests*. New York: John Wiley & Sons, Inc., 1950.
- Hamilton, C. H. "Bias and Error in Multiple Choice Tests." *Psychometrika*, 1950, 15:151-168.
- Hartmann, G. W. "Measuring Teaching Efficiency among College Instructors." *Arch. Psychol.*, N.Y., No. 154, 1933.
- Hathaway, S. R., and McKinley, J. C. "A Multiphasic Personality Schedule: I. Construction of the Schedule." *J. Psychol.*, 1940, 10:249-254.
- Hawkes, H. E., Lindquist, E. F., and Mann, C. R. *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Company, 1936.
- Jenkins, J. G. "Validity for What?" *J. Consult. Psychol.*, 1946, 10:93-98.
- Jones, R. D. "Prediction of Teaching Efficiency from Objective Measures." *J. Exp. Educ.*, 1946, 15:85-99.
- Jurgensen, C. E. "A Test for Selecting and Training Industrial Typists." *Educ. Psychol. Measmt.*, 1942, 2:409-426.

- Kitson, H. D. "Can We Predict Vocational Success?" *Occupations*, 1948, 26:539-541.
- Lawshe, C. H. *Principles of Personnel Testing*. New York: McGraw-Hill Book Company, Inc., 1948.
- Loevenger, Jane. "A Systematic Approach to the Construction and Evaluation of Tests of Ability." *Psychol. Monogr.*, 1947, 61:1-49.
- Lovell, C. "The Effect of Special Construction of Test Items on Their Factor Composition." *Psychol. Monogr.*, 1944, 56:1-26.
- Mathews, L. H. "An Item Analysis of Measures of Teaching Ability." *J. Educ. Res.*, 1940, 33:576-580.
- McAdory, M. "The Construction and Validation of an Art Test." *Teach. Coll. Contr. Educ.* 1929, No. 383.
- McPherson, M. W. "A Method of Objectively Measuring Shop Performance." *J. Appl. Psychol.*, 1945, 29 (No. 1):22-26.
- Mosier, C. I. "A Critical Examination of the Concept of Face Validity." *Educ. Psychol. Measmt.*, 1947, 7:191-206.
- and McQuitty, J. V. "Methods of Item Validation and Abacs for Item-test Correlation and Critical Ratio of Upper-lower Difference." *Psychometrika*, 1940, 5 (No. 1):57-65.
- , Meyers, M. C., and Price, Helen G. "Suggestions for the Construction of Multiple-choice Test Items." *Educ. Psychol. Measmt.*, 1945, 5 (No. 3):261-271.
- Patterson, C. H. "On the Problem of the Criterion in Prediction Studies." *J. Consult. Psychol.*, 1946, 10:277-280.
- Ruch, G. M. *The Objective or New-type Examination*. Chicago: Scott, Foresman & Company, 1929.
- Segel, D. "Construction and Interpretation of Differential Ability Patterns." *J. Exp. Educ.*, 1934, 3:203-287.
- Staff, Personnel Research Section, A.G.O. "The Army General Classification Test with Special Reference to the Construction and Standardization of Forms 1a and 1b." *J. Educ. Psychol.*, 1947, 38:385-420.
- Stalnaker, J. M. "Weighting Questions in the Essay-type Examination." *J. Educ. Psychol.*, 1938, 29:481-490.
- Stuit, D. B. *Personnel Research and Test Development in the Bureau of Naval Personnel*. Princeton, N.J.: Princeton University Press, 1947.
- . "The Effect of the Nature of the Criterion upon the Validity of Aptitude Tests." *Educ. Psychol. Measmt.*, 1947, 7:671-676.
- Symonds, P. S. "Choice of Items for a Test on the Basis of Difficulty." *J. Educ. Psychol.*, 1929, 20 (No. 7):481-493.

- Thorndike, R. L. *Personnel Selection (Test and Measurement Techniques)*. New York: John Wiley & Sons, Inc., 1949.
- Thurstone, L. L. "The Calibration of Test Items." *Amer. Psychol.*, 1947, 2 (No. 3):103-104.
- . *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947.
- University of Chicago. *Manual of Examination Methods*, 2d ed. Chicago: University of Chicago Bookstore, 1937.
- Walker, H. M. *Elementary Statistical Methods*. New York: Henry Holt and Company, Inc., 1943.
- Wesman, Alexander G. "The Usefulness of Correctly Spelled Words in a Spelling Test." *J. Educ. Psychol.*, 1947, 37:242-246.



INDEX

- Achievement tests, 9, 10, 14
- Adkins, Dorothy C., 16, 37, 131, 166
- Administration, 18-19, 21
 - cost of, 22
- Applications of test-construction principles, 2, 3
- Aptitude, 13-14
 - clerical, 182
 - and skill, 131
- Aptitude tests, 9, 10
- Arrangement, of ideas, 61, 83-85, 116
 - spiral, 32
- Battery, 15
- Bean, K. L., 90
- Cheating, 22, 111
- Civil Service Assembly of the United States and Canada, 7, 175
- Civil service examinations, 92
- Classification of tests, 8-11
- Clerical aptitude, 182
- Completion items, 9, 75-80
- Computing chart for tetrachoric correlation, 157
- Controversial statements, 51-52, 58, 61, 69-70
- Correlation, item-test, 154-155, 157-158, 164
 - product-moment, 164
- Criterion, 161-165
- Critical score, 170-171
- Criticism of examinations, 1-5
- Deciles, 174
- Dependence of items, 63-64
- Dictation, 88
- Difficulty change in test items, 7-8
- Distractors, plausible, 70-71, 82, 146
- Distribution, of scores, 125-126
 - bimodal, 170
 - skewed, 125-126
- Double negatives, 50
- Dvorak, B. J., 29
- English usage, 88-91
- Errors, in key, 148
 - typographical, 151-152
- Essay questions, 3-4, 107-108
 - advantages, 108-111
 - disadvantages, 111-116
 - examples, 116-121
 - scoring, key for, 124-127
 - objectivity in, 121-123
 - validity of, 133
- Examinations, criticism of, 1-5
 - nervousness on, 93
- Face validity, 130, 133, 160, 179
- Factor analysis, 176
- Gilbreth, F. B., 27
- Gilbreth, L. M., 27
- Grades, 5-6
- Grading (*see* Scoring)
- Graphic item counter, 155
- Greene, Edward B., 8
- Group factors, 177
- Handwriting, 112
- Ideas, arrangement of, 61, 83-85, 116
- Illiterates, employment tests for, 91-92, 94-95
- Instructions, 18-19, 21, 94-95
 - to consultants, 146
- Intelligence, 10, 12, 13, 103-104
 - general factor of, 176-177
 - validity of tests of, 162
- Internal consistency (*see* Correlation, item-test)
- Irrelevant information, 66-68, 73
- Items, 15
 - analysis of, 90, 152-158
 - completion, 9, 75-80
 - construction of, rules for, 37
 - dependence of, 63-64
 - matching, 9, 80-83
 - multiple-choice (*see* Multiple-choice items)

- Items, performance, 9
 - short-answer and simple recall, 9, 75-80
 - true-false (*see* True-false items)
 - validity of, 153-154
- Job analysis, 27-30, 40
- Length, 21
- Louisiana Department of State Civil Service, 175, 179
- McQuitty, John V., 157
- Matching items, 9, 80-83
- Material, security of, 145
 - sources of (*see* Sources of material)
- Mosier, Charles I., 55, 157
- Multiple-choice items, 9, 52
 - distractors, 53
 - examples, 55-62
 - for illiterates, 92
 - instructions, 54
 - punctuation, 52
 - review, 147
 - rules, 62-75
 - vocabulary, 105
- Myers, M. Claire, 55
- Nervousness on examinations, 93
- Norms, 169-171, 173-175
- Objective questions, 3-4, 9
- Outline, of subject matter, 85-87, 148
 - of test, 31-36, 181
- Percentiles, 173
- Performance items, 9
- Performance tests, 129-131
 - for counselors, 134-136
 - material and procedure, 131-134
 - and responses of counselors, 189-222
 - scoring, 139-141
 - validity of, 113
 - what they measure, 138
- Personnel selection, 6, 17-18, 88-96, 130, 179-187
- Plausible distractors, 70-71, 82, 146
- Premise and choices, 59
 - clear problem in, 65-66
- Price, Helen G., 55
- Projective techniques, 108
- Rating scales, 162-163
- Reading comprehension, 96-101
- Reliability, 15
 - definition of, 161
 - of English usage test, 91
 - of essay tests, 113-114
 - factors determining, 165-166
 - Kuder-Richardson formulas, 168
 - of performance tests, 132
 - split-half method, 167-168
 - test-retest method, 167
- Research, 176
- Review, from construction viewpoint, 150-152
 - by consultants, 143-150
 - of sample material, 184
 - by writer, 142
- Rinsland, Henry Daniel, 124
- Rogers, Carl, 188
- Ross, C. C., 37, 46
- Rote memory, 84
- Ruch, F. L., 39, 62
- Ruch, G. M., 37
- Rules for item construction, 37
- Scoring, 10, 22-23
 - correction for chance, 25-26
 - and grading, 124
 - machine, 23-24
 - objectivity in essay, 121-123
 - for organization, 123-124
 - of performance tests, 139-141
 - use of key, 126-127
- Security of material, 145
- Sequence in answers, 49, 64, 151
- Shartle, C. L., 28
- Short-answer items, 9, 75-80
- Simple recall items, 9, 75-80
- Skewed distribution, 125-126
- Skills, 128-129, 131
- Sources of material, 38
 - Civil Service Assembly, 42
 - class discussion, 39-40
 - experts, 41
 - job analysis, 40
 - lecture, 39

- Sources of material, standardized tests,
47-48
textbooks, 39
Specific determiners, 49, 61-63
Speed test, 19-20
Spiral arrangement, 32
Standard scores, 171
Standardized tests, 10, 174-176
Statistics, 3, 159
Stone, D. R., 222
Subtest, 15
Super, Donald E., 8, 28

Tables, interpretation of, 101-102
Teacher-made tests, 10
Teaching objectives, 26-27
Techniques, projective, 108
Terman, L. M., 51
Terms, definition of, in items, 49
of the job, 147
Tests, achievement, 9, 10, 14
aptitude, 9, 10
classification of, 8-11
definition of, 11-12
employment, for illiterates, 91-92, 94-95
performance (*see* Performance tests)
speed, 19-20
standardized, 10, 174-176
teacher-made, 10
validity of, 160-165

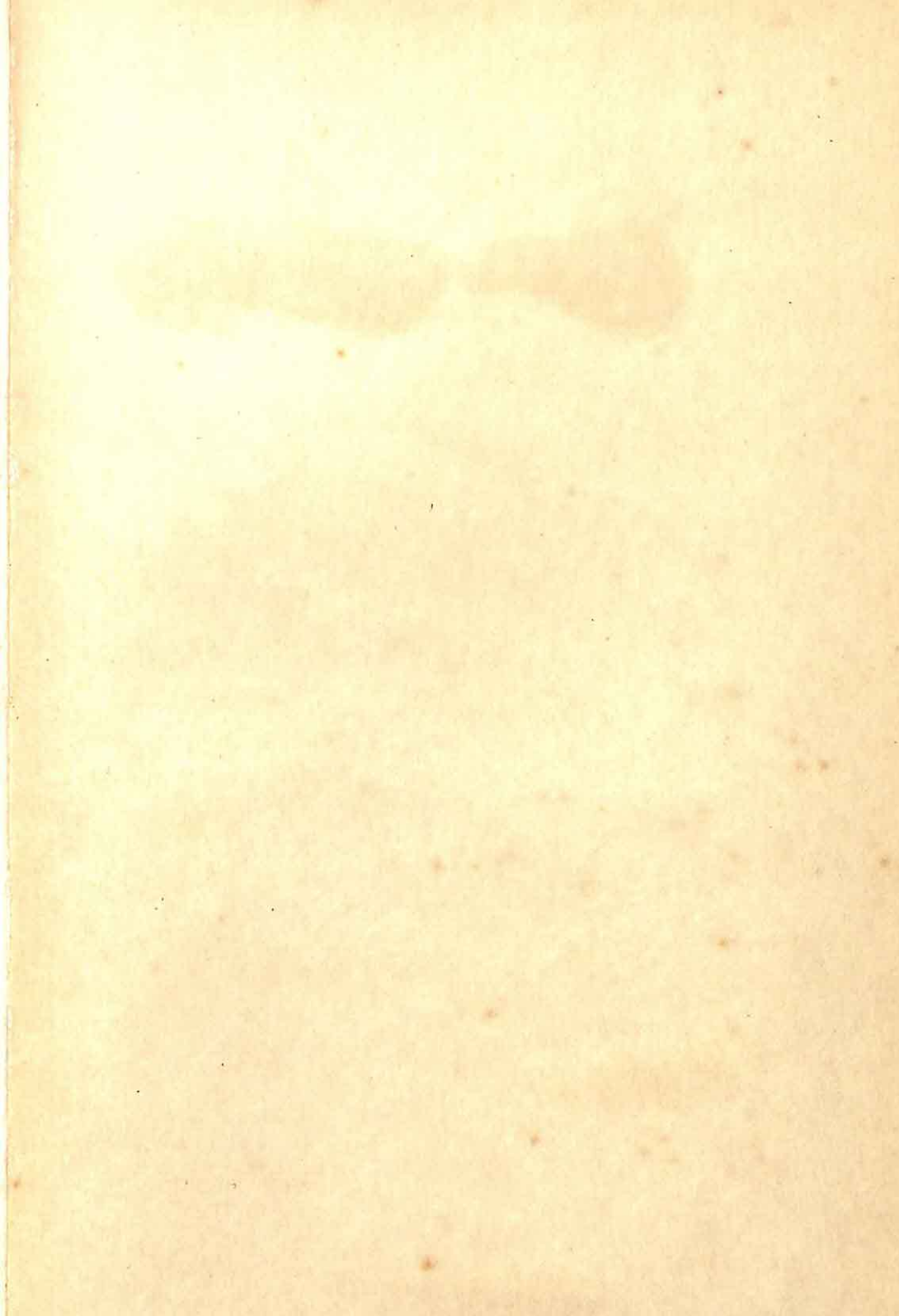
Thorndike, Robert L., 129, 131
Thurstone, L. L., 178
Trades, measurement of aptitudes and skills, 144
Trites, H. L., 188
True-false items, 9, 44
advantages, 44-45
disadvantages, 45-46
for illiterates, 92
instructions, 46
review, 147
rules, 47-52
zero, 100-101

Unique factors, 177

Validity, 15, 30-31
of essay questions, 113
face, 130, 133, 160, 179
of items, 153-154
of performance tests, 133
of tests, 160-165
Viteles, M. S., 27
Vocabulary in test items, 72-73, 103-106, 182

Wechsler, David, 51, 173
Weighting, 24-25
Work sample, 9, 18, 129
(*See also* Performance tests)

Yes-no form, 43-44, 92, 94-95



Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological
Research Library.**

The book is to be returned within
the date stamped last.

17. 6. 6

18. 6. 6

23. 6. 6

17 JUN 1965

20 JUN 1965

- 6 JUN 1967

371.26
BEA
Form No. 4

BOOK CARD

Coll. No.

Accn. No. 591

Author.. Bean, ...

Title.. Construction of Educ. Perso.
nal test

Date.

Issued to

Returned on

17.6.6 7060
17 JUN 1965

